# Information Visualization Evaluation Using Crowdsourcing

EUROGRAPHICS 2018 / J. Heer, H. Leitte, and T. Ropinski
(Guest Editors)

*Volume 37 (2018), Number 3*

# Information Visualization Evaluation Using Crowdsourcing

R. Borgo[1*], L. Micallef[2*], B. Bach[3], F. McGee[4], B. Lee[5]

[1]Informatics Department, Kings's College London, United Kingdom
[2]Department of Computer Science, Aalto University, Finland
[3]School of Informatics, University of Edinburgh, United Kingdom
[4]Environmental Informatics Unit, Luxembourg Institute of Science and Technology (LIST), Luxembourg
[5]Microsoft Research, USA

**Abstract**

*Visualization researchers have been increasingly leveraging crowdsourcing approaches to overcome a number of limitations of controlled laboratory experiments, including small participant sample sizes and narrow demographic backgrounds of study participants. However, as a community, we have little understanding on when, where, and how researchers use crowdsourcing approaches for visualization research. In this paper, we review the use of crowdsourcing for evaluation in visualization research. We analyzed 190 crowdsourcing experiments, reported in 82 papers that were published in major visualization conferences and journals between 2006 and 2017. We tagged each experiment along 36 dimensions that we identified for crowdsourcing experiments. We grouped our dimensions into six important aspects: study design & procedure, task type, participants, measures & metrics, quality assurance, and reproducibility. We report on the main findings of our review and discuss challenges and opportunities for improvements in conducting crowdsourcing studies for visualization research.*

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation

## 1. Introduction

Crowdsourcing can overcome a number of limitations of controlled laboratory experiments, including small participant sample sizes and narrow demographic backgrounds of study participants. As these benefits can further improve important aspects of empirical research, such as generalization of empirical results and their ecological validity, the visualization research community has been actively employing crowdsourcing for conducting empirical research. On the other hand, these benefits of crowdsourcing accompany conceptual and methodological challenges for rigorous empirical visualization research.

To shed light on this emerging phenomenon, a seminar entitled "*Evaluation in the Crowd: Crowdsourcing and Human-Centred Experiments*" was held in November 2015 as part of the Dagstuhl Seminars series[†]. Researchers from the human computer interaction and information visualization communities, as well as researchers investigating crowdsourcing practices were gathered to share experiences in conducting empirical research. They also discussed limitations and methodological considerations in designing and deploying crowdsourcing experiments aimed at evaluating effectiveness of visual representations. The main outcome of the seminar was a book with the same title as the seminar [APH17]. The book provides a high level review of concerns and discussions with respect to crowdsourcing and human-centred experimentation methodologies from different perspectives.

During the seminar, the authors of this article (along with other seminar participants) worked on a book chapter titled "Crowdsourcing for Information Visualization: Promises and Pitfalls" [BLB*17], looking at crowdsourcing specifically from the visualization point of view. In the book chapter, we discussed core aspects for successful employment of crowdsourcing in empirical studies for visualization – participants, study design, study procedure, data, tasks, and metrics & measures – reflecting on our own experiences of running crowdsourcing experiments and a set of selective papers that employed crowdsourcing for visualization research. We also discussed potential ways to overcome the common pitfalls of crowdsourcing, and provided four case studies as good examples.

In this paper, we present a deeper, more focused, and more systematic review of existing literature that has employed crowdsourcing for empirical evaluations of visualizations. Our goal was to capture the practices of crowdsourcing evaluations for visualization in terms of how researchers designed and reported crowdsourcing studies. We report on the designs, methods, tasks, tools, and measures used in crowdsourcing studies over the decade. Our analysis has produced very interesting findings. For example, we confirmed that the number of papers employing crowdsourcing experiments is following an upward trend, but to our surprise the first paper with a crowdsourcing experiment was published in 2009 (Figure 1). While

* First and second author share equal contribution to the work and are presented in no particular order.

† http://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=15481

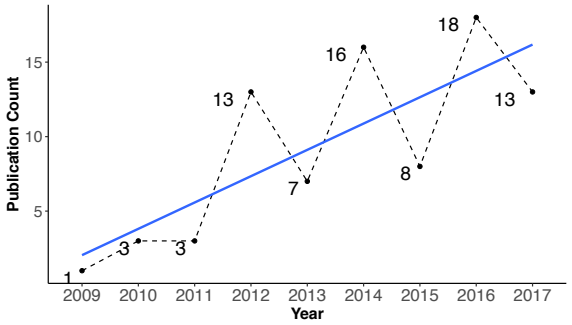


**Figure 1:** *The number of papers employing crowdsourcing experiments has followed an upward annual trend. While we analyzed papers published between 2006 and 2017, the first paper with a crowdsourcing experiment for visualization was published in 2009.*

it is often argued that time is not a reliable measure in crowdsourcing experiments, time and error are the most popular measures used in crowdsourcing studies. Despite the awareness regarding the inherent limited control over participants selection and study settings, only a limited percentage of experiments reported to have controlled the worker's performance profile (39%) or included checks to identify and filter out inattentive participants (11%). Comparatively the most surprising finding was that, despite the importance given to rigour in empirical evaluation in information visualization, many papers failed to properly report important details about the study, making it difficult not only to reproduce the study, but also to even assess the validity of the study.

We also expand on potential pitfalls through the discussion of observed issues and mechanisms to overcome them, and discuss the evolution of crowdsourcing experiments now playing a crucial role in the field of empirical studies in information visualization. This report is a first step toward enabling the collection of high quality empirical data, reproducible user study findings, and improved visualization techniques. Furthermore, it provides guidance to researchers in appropriately reporting all the necessary details about their crowdsourcing experiments to demonstrate data collection validity. Anyone new to crowdsourcing will benefit from this report, as it provides the means to quickly learn how to effectively use crowdsourcing for information visualization evaluation.

## 2. Background

Different aspects of crowdsourcing have been studied across research communities to, for instance, identify best practices and techniques to improve the quality of the collected data (e.g., [KKMF13, HHR*14, HF13]) or to determine effects of factors such as incentives over measured dependent variables (e.g., [SM13, SHC11]). Crowdsourcing became widely accepted for information visualization research following Heer and Bostock's seminal work [HB10] and it has since been growing in popularity, hand in hand with traditional empirical evaluation approaches.

### 2.1. Evaluation in Information Visualization

The visualization research community recognizes the importance of evaluation. Common and unique challenges in the use of empirical evaluation for information visualization research have often been discussed in our community, particularly at the biennial BELIV (evaluation and BEyond - methodoLogIcal approaches for Visualization; formerly, Evaluation and Beyond - Methodological Approaches for Visualization) workshop series. Plaisant summarized evaluation practices and challenges specific to information visualization, and proposed steps for improvement [Pla04] comprising of: the development of (data and task) benchmarks and repositories; the dissemination of success stories; and the refinement of toolkits for adoption. Robertson et al. emphasized the importance of utility and usability testing [RCFL09]. In contrast, Lam et al. classified evaluation methods based on the goals of the visualization research [LBI*12], including the evaluation of: environments and work practices; visual data analytic and reasoning; communication through visualization; collaborative data analysis; user performance; and user experience, visualization algorithms.

Another line of discussion was centered on broadening the evaluation methodologies for visualization evaluation. For example, Carpendale emphasized the importance of empirical research and encouraged careful application of broader types of empirical methodologies in information visualization [Car08]. Brehmer et al. called for reflecting upon the unique challenges associated with visualization design, which are confronted during the use of empirical methods at early stages of design [BCLT14]. They noted the importance of characterizing work practices and associated problems in a specific domain to motivate design choices during the later development of a specific application or technique. More recently, Thudt et al. encouraged the adoption of evaluation methods that go beyond measuring performance via controlled experiments in the context of personal visualization [TLCC17].

### 2.2. Controlled Experiments in Information Visualization

Controlled experiments are commonly used to evaluate the effectiveness of graphical encodings, visualization techniques, and visual analytics systems. Effectiveness is typically measured based on user performance (task-completion time, error rate, number of clicks) and user experience (perceived task difficulty, subjective preferences, difficulty in using the visualization). Aspects such as the user tasks (e.g., task difficulty) and data characteristics (e.g., data complexity) are typically controlled to avoid effects from compounding factors. Experiments are usually carried out in a laboratory for full control over the environment and system specifications. In some cases, special equipment for interaction (e.g., touch displays, tabletop displays, display walls, virtual reality) or recording devices (e.g., microphones, cameras, eye-tracking devices) are specifically required to achieve the experiments' goals. The experimenters can closely observe the participants' execution and directly interact with them.

However, controlled environments do not adequately capture the realistic scenario of in-the-wild usage. They can also be very costly as the participants need to be compensated adequately for their time to come to the laboratory. The higher the cost, the fewer partici-

pants can be recruited. Nonetheless, a relatively large pool of participants is required to collect enough data for statistical analysis. All of these factors have contributed to the proliferation of the use of crowdsourcing in conducting empirical research across many domains, including information visualization.

### 2.3. Crowdsourcing

The term *crowdsourcing* was first introduced in 2006 [How06] to describe a new marketplace whereby anyone having Internet access (also known as the *crowd*) could contribute to the completion of a task for free or in exchange for a monetary reward. Over the years, crowdsourcing became a well-accepted marketplace for recruiting large groups of participants for empirical research in diverse disciplines. This led to the era of *crowd science* and the definition of *passive* versus *active* crowdsourcing [KSW15].

Active crowdsourcing is identified as internet users (also known as *crowdworkers*) contributing to a specific solicited task using a software platform, such as Amazon Mechanical Turk (AMT) or Crowdflower (recently been rebranded as "Figure Eight") . In contrast, passive crowdsourcing is when publicly available information (e.g., on web sites or online social platforms) is collected without any solicitation. Human-Intelligence-Tasks (HITs, also called *microtasks*) are core to active crowdsourcing. HITs are small tasks that typically do not require any specific or prior knowledge to be completed. Crowdworkers who complete a HIT should be paid, but often are not paid well enough to obtain a living wage, raising critical ethical and socio-economical issues [MCG*17].

### 2.4. Crowdsourcing versus Controlled Experiments

Scientific research typically adopts active crowdsourcing. Crowdsourcing platforms support the experimenter in aspects such as organization, management, and delivery of HITs. Such support includes access to a (registered) pool of crowdworkers, which eases the recruitment process and opens up opportunities to scale-up the size of the participant pool for empirical studies. Platforms also provide a web infrastructure for hosting, advertising, and completing HITs. Additionally, the centrally managed system handles the supply/demand and automatic payments of the crowdworkers, and manages communication between crowdworkers and providers. They also provide elementary measures to control and ensure the quality of the collected data. For instance, AMT evaluates the quality of the tasks completed by crowdworkers using a HIT approval rate (i.e., the number of completed HITs approved as by providers as of good quality to the total number of completed HITs). Such measures have made crowdsourcing a prominent and well-accepted recruitment mechanism within various research communities, including visualization, HCI, machine learning, and other related fields. Yet, technical limitations of crowdsourcing platforms [HJR*17] impose a number of challenges and constraints on experiment design [GMN*17, BLB*17].

Through crowdsourcing, larger participant pools can be easily recruited faster and at a much lower cost than in controlled laboratory experiments. The more participants, the larger the sample sets and the greater the chance to attain statistical significance. The participating crowdworkers are likely to have very diverse backgrounds and demographics, thus ensuring the generalizability of the findings derived from the user experiments. One study that empirically compared different visualizations for Bayesian reasoning with the classic representations of the problem, could not replicate the results of previous controlled laboratory studies. Their participants' accuracy was exceptionally lower, despite that they checked for and filtered out responses from inattentive participants [MDF12]. The authors argued that participants in the laboratory, particularly if they are university students, are less likely to adequately represent the mind and skill set of the general population and the real-world scenario where timely but accurate decisions are made. Workers want to complete tasks quickly but also accurate to example improve their HIT approval rate. Thus, crowdsourcing allows us to better evaluate our visualizations in-the-wild, particularly when our visualizations are aimed to support the general population with no specific background or skill set. Having access to a diverse population also helps experimenters find workers with specific domain knowledge or skills. However, in such cases, specific tests and checks need to be devised to ensure that only the workers that satisfy the experiment's requisites are qualified and allow to participate in the study.

Crowdworkers are anonymous, the environment where they complete the task is not controlled, and the experimenter cannot directly interact or observe participants. Such factors can have undesired impacts on the user study and the quality of the collected data. Thus, in crowdsourcing experiments, the experimenters have to devise and include clever quality assurance measures that are typically not required in controlled laboratory experiments. For instance, the online experiment should ensure that both the hardware (e.g., devices used) and the software (e.g., the version of the web browser) used by the participant adhere to the experiment's system requirements. Specialized questions and measures need to be crafted to assess the participant's attention while completing the task. A more comprehensive comparison of crowdsourcing versus laboratory experiments, highlighting their pros and cons and when and when not to use each one, is provided in Gadiraju et al.'s recent book chapter [GMN*17].

### 2.5. Crowdsourcing for Information Visualization

In 2008, van Ham and Rogowitz [vHR08] were the first to conduct online experiments for Information Visualization research. Their experiment collected data on how users lay out graphs to best represent the structure in the data. As they did not use crowdsourcing for evaluation, we excluded this paper from our survey. However, this seminal work demonstrated the opportunities (and partly the challenges) crowdsourcing has to offer to the visualization community. Since then, the use of crowdsourcing for visualization research has proliferated at a very steady rate and as years pass by, more sophisticated crowdsourcing experiments are being conducted using this paradigm (see Section 4).

Different core aspects that are critical to the success of a crowdsourcing experiment for visualization research have recently been identified and discussed by prominent researchers in the field in Borgo et al.'s book chapter [BLB*17]. These aspects consist of (i) participant selection, (ii) study design, (iii) study procedure, (iv)

| | |
|---|---|
| **Participants** | Anonymous crowdworkers with diverse cultural backgrounds, demographics, skills, traits (e.g., colorblindness) and abilities (e.g., visual literacy) that might not match the experiment's requirements. Also, crowdworkers typically aim to complete several tasks, at times in parallel with less amount of effort and as quickly as possible. |
| **Study Design** | Crowdsourcing platforms do not provide support for preventing multiple participation. Yet, between-subjects design is often preferred for crowdsourcing experiments to ensure that the experiment is short enough for the participant to remained focused, and to have full control of extraneous variables (e.g., show only one of the assessed visualizations). |
| **Study Procedure** | Experimenters cannot directly interact with the crowdworkers (e.g., to train workers how to interpret a new visualization technique) or to directly observe the participant's work (e.g., to infer insights on the reasoning process). The working environment cannot be controlled, yet certain complex visualization experiments require focused attention. Special equipment (e.g., touch displays), which at times are critical for the visualization experiment, cannot be used. The expected system requirements have to checked remotely (e.g., to ensure that all participants can view and interact with the visualization as expected). |
| **Study Data** | Controlled data especially synthetic data that does not have a real-world context (common in visualization studies) might be difficult for most crowdworker to adhere to and to complete the task effectively. Yet certain real-world data, especially sensitive data (e.g., patient data for a visual analytics healthcare system), cannot be used due to data privacy and confidentiality. |
| **Study Tasks** | Visualization studies at times involve tasks (e.g., collaborative visual analytics) that require more time and attention than is typically recommended and possible for a crowdsourcing microtask. Adequate amount of time should be specifically dedicated for clear instructions, training, abilities tests, attention checks and questionnaires (e.g., for demographics since workers are diverse and anonymous), besides the actual task. |
| **Study Measures** | User performance measures commonly used in visualization, such as time, might not be entirely reliable due to, for example, hardware latency. Similarly for error if example workers use unexpected helping aids. Measures that take into account,for example eye movement (e.g., to study visual saliency), cannot be used. Additionally, quality assurance measures devised to monitor attention have to be adequately designed for the experiment. |

**Table 1:** *A summary of the critical challenges specific to crowdsourcing Information Visualization experiments with respect to the six aspects identified in Borgo et al.'s book chapter [BLB*17].*

study data, (v) study tasks, and (vi) metrics and measures. The identified aspects are also critical for the success of a controlled laboratory study. However, there are a number of challenges that are specifically critical for crowdsourcing experiments with respect to each aspect (see Table 1 for summary and Borgo et al.'s book chapter for a detailed discussion). Moreover, aspects like measures for quality assurance are particularly important and critical for crowdsourcing experiments since, as discussed earlier, the environment is not entirely controlled and the experimenter does not directly interact with the participants. Borgo et al.'s book chapter also provides a detailed review of the opportunities and challenges that crowdsourcing brings to visualization evaluation, when and when not to use crowdsourcing, and case studies showing the diverse use of crowdsourcing for visualization research.

## 3. Methodology

Our focus was on information visualization research that used crowdsourcing for evaluation. Thus, we selected papers from major academic information visualization outlets, such as conferences and journals, between January 2006 and December 2017 (Table 2). We used the digital library of each individual outlet to search for all publications whose title, abstract, or the actual manuscript satisfied the following query: (*MTurk* `or` *crowdsourced* `or` *crowdsourcing* `or` *crowd* `or` *turk*) `and` ( *visualization* `or` *visualisation*).

To determine a useful set of dimensions, each author of this survey performed an initial analysis of 20 papers. We first focused on the six aspects Borgo et al. [BLB*17] identified as the core of a crowdsourcing visualization experiment (see Section 2.5). We then discussed our initial analysis with respect to Borgo et al.'s aspects, and identified a set of the dimensions that would concretize important factors within each aspect that could affect the success of a crowdsourcing visualization experiment. We later performed a second round of analysis using our set of dimensions. We divided the remaining set of papers, and individually performed an analysis with the subset of papers. We had a series of semi-regular meetings over Skype to discuss potential papers for exclusion and to fine-tune the set of dimensions, when necessary. We excluded the papers that satisfied one or more of the following criteria:

- used crowdsourcing only for the pilot study (e.g., [TGH12]);
- used crowdsourcing not for evaluation but for other purposes such as data collection (e.g., user generated layouts [KDMW16, KWD14, vHR08]);
- evaluated graphics but no information visualization (e.g., [BCER14, XADR13, KKL16]);
- evaluated user interfaces but no information visualization (e.g., [DCS*17, MBB*11, RYM*13]);
- evaluated human-computer interaction but no information visualization (e.g., [MMS*08, KWS*14, DH08]);
- proposed a novel crowdsourcing platform (e.g., [TBRA17, EKR16]);
- used already available data maintained by a crowdsourcing platform without using crowdsourcing (e.g., [KHA16, DDW11]);

| Outlet | # of Papers (# of Experiments) |
|---|---|
| IEEE Transactions on Visualization and Computer Graphics (TVCG) + IEEE Information Visualization (InfoVis) + IEEE Visual Analytics Science and Technology (VAST) | 37 (86) |
| ACM Conference on Human Factors in Information Systems (CHI), including Extended Abstracts | 24 (59) |
| Computer Graphics Forum (CGF) + EG\VGTC Conference on Visualization (EuroVis), including Short Papers | 16 (24) |
| ACM Transactions on Computer-Human Interaction (TOCHI) | 4 (19) |
| IEEE Pacific Visualization Symposium (PacificVis) | 1 (1) |
| International Conference on Advanced Visual Interfaces (AVI) | 1 (1) |
| Total | 82 (190) |

**Table 2:** *The number of surveyed papers and experiments published in different journals and conferences between 2006 and 2017.*

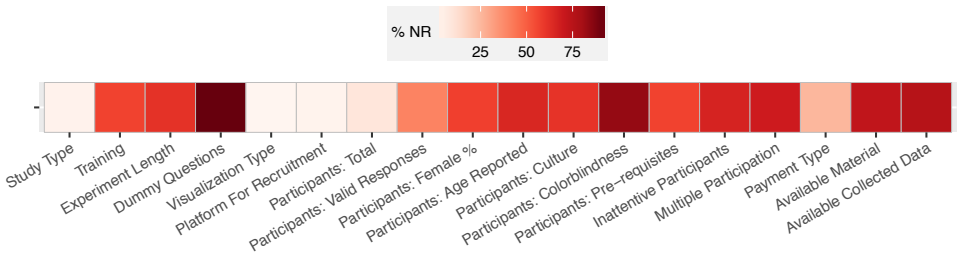


**Figure 2:** *A heatmap showing the proportion of experiments for which no value was recorded for the specified dimensions.*

- discussed previous crowdsourcing studies without using crowdsourcing (e.g., [AJB16, ZGB*17, KH16]);
- mentioned the use of crowdsourcing for future evaluation (e.g., [HLS16]).

Our analysis resulted in a final set of 82 papers[‡], containing a total of 190 experiments. We concluded with the 36 dimensions listed in Table 3, categorized into the six aspects discussed in our review: study design & procedure, task types, participants, measures & metrics, quality assurance, and reproduciblity. Certain dimensions might not be applicable to all experiments. For example, colorblindness is not important if the evaluated visualizations do not use color. However, we wanted our dimensions to capture all factors that could affect any visualization experiments using crowdsourcing. We omitted the data aspect from Borgo et al. as the details we provided for the task type partly covered this aspect with respect to task versus data type. Borgo et al. also address orthogonal data aspects, such as attractiveness, as a means to increase participants engagement and task appeal, which we discuss in Section 6.4 when we describe motivators. Other data aspects such as familiarity and confidentiality are fully addressed by Borgo et al. We, however, added two new aspects, quality assurance and reproducibility, as these aspects required more attention. While quality assurance is fundamental for ensuring valid and reliable empirical data, reproducibility of our experiments is helpful for the future of any research community [KLB*16], including InfoVis. We did not include the objectives of experiments in our dimensions because all of our surveyed visualization experiments used crowdsourcing for evaluation. We also excluded the research questions handled by the experiments because such questions are typically very specific to the context of the handled problem, the comparable small size of literature also did not heed yet any interesting pattern with respect to crowdsourcing evaluation.

Our analysis revealed that many papers are not free from reporting issues, and fail to report important details about the experiment (Figure 2). We tagged such experiments as not having reported details about the dimension, instead of assuming that that dimension was not considered. As missing details make it difficult not only to reproduce experiments but also to assess them, we propose ways to alleviate these reporting issues in Section 5.1.

## 4. Findings

Crowdsourcing became widely accepted for information visualization (InfoVis) research following Heer and Bostock's 2010 seminal work, which compared graphical perception experiments conducted in the laboratory with similar experiments conducted using AMT [HB10]. The number of papers reporting crowdsourcing experiments for InfoVis changed from one in 2009 to 13 in 2017 (Figure 1). While papers in the early rise of crowdsourcing had to dedicate an entire section to justify the use of crowdsourcing for their empirical studies (e.g., [MDF12]), recent papers simply cite previous InfoVis papers to justify the use of crowdsourcing for evaluating data visualizations (e.g., [MPOW17]).

[‡] The full list of the surveyed papers tagged along all dimensions can be found online at https://crowdsourcing4vis.github.io.

| Aspect | Dimension | | Definition | Tag |
|---|---|---|---|---|
| Study Design & Procedure | Study Type | Between | follows between-subjects design | yes, no |
| | | Within | follows within-subjects design | yes, no |
| | | Mixed | design not uniquely classifiable as between- or within-subjects | yes, no |
| | Experiment Length | | length of experiment in minutes | number |
| | Training Type | | passive (e.g., reading tutorial) or active (e.g., answering test-like questions) | passive, active |
| | Dummy Questions | | test questions (e.g., to distract users) not part of training and not analyzed | yes, no |
| | Visualization Type | Static | task stimuli were static | yes, no |
| | | Interactive | task involved interaction with stimuli | yes, no |
| | Recruitment Platform | | platform used to recruit participants | AMT, CrowdFlower, Other |
| | Device & Software Restrictions | | type of device and software restrictions enforced for participation | string |
| Task Types | Task Type | | type of task completed by the participants in the experiment | string |
| Participants | Number | Total | total number of participants that took the experiment | number |
| | | Per Condition | number of participants per condition, if applicable | number |
| | | Valid Responses | total number of participants with valid responses | number |
| | Gender | Female % | percentage of participants that were female | number |
| | Age | Reported | age was reported | yes, no |
| | | Range | age range | [number,number] |
| | | Mean | age mean | number |
| | Culture | USA Only | participants were all from the USA | yes, no |
| | | USA:India:Europe:Other | percentage of participants per mentioned continent | number:number:number:number |
| | Requisites | Colorblindness | how was colorblindness controlled for | self-reported, tested |
| | | Prerequisites | requirements, determined a priori (no"aptitude test"), for participation | string |
| | | Pre-test | any "aptitude test", not determined a priori, for participation | string |
| Measures & Metrics | Error | | error (absolute or relative) was a measured and analyzed | yes, no |
| | Time | | time measured and analysed | yes, no |
| | Confidence | | confidence in participant's response measured and analyzed | yes, no |
| | Abilities | Numeracy | participants' numeracy ability measured and analyzed | yes, no |
| | | Spatial | participants' spatial ability measured and analyzed | yes, no |
| | Other | | any other dependent variables measured and analyzed | string |
| Quality Assurance | Inattentive Participants | | strategy to detect inattentive participants | Catch questions task relevant, Catch questions task irrelevant, Task completion time threshold, Other |
| | Multiple Participation | | strategy to ensure that workers took the experiment only once | pre study, post study |
| | Payment | Type | if payment was per trial, per study or any other payment criteria | per trial, per correct trial, per study, raffle, other |
| | | Amount | payment participants received | number |
| | Bonus | | type of bonus granted | string |
| Reproducibility | Available | Material | type of publicly available material, e.g., code, stimuli, questions | string |
| | | Collected Data | collected, anonymized data is publicly available | yes, no |

**Table 3:** *Our 36 dimensions categorized into six aspects by which we tagged and analysed 190 experiments. Papers that did not report about the dimension were tagged with NR (Not Reported).*

In this section, we report on the findings from our analysis of the 190 experiments with respect to the aspects and dimensions discussed in Section 3.

### 4.1. Study Design and Procedure

The design phase of crowdsourcing experiments is more complex than that of controlled laboratory experiments [GMN*17]. While crowdsourcing facilitates participants recruitment, it still relies heavily on the experimenters' expertise to determine if a study is suitable to be crowdsourced and to carry out quality control on both participant selections and data collection. The latter are challenged by the lack of control over participants behavior and experimental environment.

**Study Type**: With between-subjects experiments, each participant goes through only one of the conditions being tested while with within-subjects experiments, each participant completes all conditions. The mixed design reflects experiments where multiple independent sets of conditions may have been evaluated and participants may have seen all of one and not the other. Across the 190 experiments covered by our survey, 40% were between-subjects design, 40% were within-subjects design, 17% were a mixed design, and 3% were unspecified. The number of publications using each study type by year is shown in Figure 3.

**Experiment Length**: Only 37% of the experiments specify an approximate average duration for the experiments. The average duration was 15.6 minutes and the maximum and minimum were 125 and 0.3 minutes, respectively.

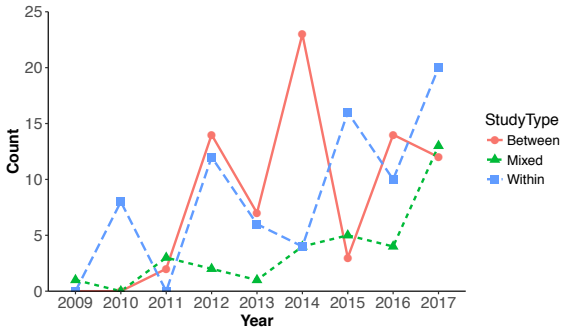


**Figure 3:** *The number of publications by study type, per year.*

**Training:** While training is important for ensuring that a participant fully understands a task, only 43% of experiments reported training their participants. Of those experiments, slightly over two thirds (71%, 31% of the total) employed active training, where the participant actually undertook training trials. The rest (30%, 13% of the total) employed passive training, where the participants training was in the form of non-interactive instructions and examples.

**Dummy Questions:** Along with randomizing the order of questions, dummy questions (also known as throw away questions) are a useful technique in mitigating the learning effect for user experiments [Pur12, chapter 3]. The idea is to not count an initial set of trials where the participant may be less sure of the task or the experiment interface as they are still learning how to complete the task. Only 6% of the papers reviewed (15 experiments total) specified using dummy questions. For crowdsourcing based experiments, it can be argued that training and dummy questions are more important than in lab experiments, as a user cannot, usually, contact the experimenter if they are not sure of the task. It is not clear whether more experimenters actually did use some form of training and dummy questions but did not report it due to space restrictions. The experiments that reported using dummy questions were of both between-participants and within-participants, even though the learning effect is more strongly associated with within-participants experiments.

**Recruitment Platform:** Of all the papers we reviewed, 82% reported using AMT as their platform of choice (e.g., [OPH*16, TLM16, ZK09]) and 6% did not specify the platform that was used (e.g., [BDF15, HCL12]). The remaining 12% used one of the following: CrowdFlower (e.g., [VC17, CAB*16]), MicroWorkers (e.g., [SBJS15]), Reddit (e.g., [KKHM16]), and self-deployed web application (e.g., [BNK*16]) or game (e.g., [AZM12]). The inclusion of participant prerequisites could be one of the reasons why AMT is so popular. The use of the platforms other than AMT has begun to increase over the past four years (up to four papers in 2016 and 2017), probably because AMT has been restricted to only USA use since then. However, AMT is still dominant.

**Visualization Type:** An important design choice when building a visualization is whether it will be a static or interactive visualization. The majority of the experiments (74%) used a static visualization approach, where there was no interaction. This is slightly surprising due to the ability of modern web development tools for visualization to easily incorporate interactive aspects. One key factor that may have impacted this is the ease of deployment on AMT. As mentioned above, AMT is the most popular platform. However, it is primarily designed for displaying images and text to participants. To include interactivity requires the experimenter to build and host the experiment elsewhere and serve it to the AMT page in an embedded frame. Additionally for some experiments interactivity may be considered a confounding factor. However, as modern visualizations are rarely purely static, depending on context, evaluating them with some form of realistic interaction may be considered more valid.

**Device & Software Restrictions**: A final consideration in study design is the use of a specific experiment setup. Lab studies often require specific experimental software and/or hardware. While not frequent or as broad in range as for lab studies, crowdsourcing based studies sometimes require specific testing setups (e.g., minimal display sizes, screen resolution, higher end interaction capabilities [KHD12, FSASK13]) and data collection through specialized I/O devices, for example eye tracking. Such restrictions while not common (only 15% of papers) are possible. Within our survey some authors have restricted based on device type (e.g., [SOK*16,BEDF16,KKHM16,BKH14,HRC15]), using the options offered by AMT. With respect to eye-tracking, modern commodity webcams can offer a workaround to using specialist hardware, see [XEZ*15, LHM*15], and an alternative approaches to using a camera is possible [KBB*15].

### 4.2. Task Types

Tasks are a building block of any user study and as such needs careful design. Crowdsourcing settings introduce different challenges when dealing with task design in terms of semantics of the task (e.g., cognitive complexity and skills requirements), platform constraints (or lack of), and response time. Some tasks require short response times (e.g., [KLKY11]) and others require specific skills and more complex reasoning (e.g., [CTIB15]).

The characteristics of a task can be classified according to a taxonomy. Shneiderman [Shn96] suggests a task taxonomy by data type, listing several low-level tasks required to perform analysis and exploration: overview, zooming, filtering, details-on-demand, relation finding, historical tracking of actions, and extraction. Ward et al. [WGK10] identify exploration, analysis, and presentation as three abstract tasks a user seeks to accomplish with visualization. In addition, Andrienko and Andrienko [AA05] and McEachren [Mac04] proposed other models, focusing more on time varying data. *Elementary tasks* (from [AA05]) consist of data element look-up, comparison, and relation seeking. *Synoptic tasks* involve patterns and trends in the data and relationships within data or to external data. While Schneiderman's taxonomy is more detailed on the type of action performed during analysis, Andrienko and Andrienko's classification better captures the type of tasks typical of a crowdsourcing experiment. Thus, we categorized tasks based on Andrienko and Andrienko's classification:

**Elementary tasks**

- outlier detection (e.g., [CMFH12, ACG14, MPOW17]);
- qualitative comparison (e.g., [SSG16]);
- qualitative estimation (e.g., [DBH14]);
- quantitative comparison (e.g., [KHA10]);
- quantitative estimation (e.g., [SK16, KS16, HAS11]);
- target identification (e.g., [ZK10, FFB18]).

**Synoptic tasks**, requiring an holistic view of the data

- decision making (e.g., [HDR*13]);
- pattern and trend detection (e.g., [ZRH12]);
- clustering (e.g., [MPOW17]);
- data filtering on multiple criteria (e.g., [JS10, GCNF13]);
- visual interpretation (e.g., [IFBB12]);
- Bayesian inference, uncertainty and likelihood estimation (e.g., [MDF12, OPH*16]).

**Mixed tasks** involve both exploration and analysis (e.g., [BKH14, RSC15, BEDF16]).

Profiling the literature, a pattern emerges (Figure 4), which indicates elementary tasks as being favoured with respect to synoptic. Overall 60% of the studies implemented elementary tasks, 30% synoptic, and 10% mixed, involving both task typologies within the same trial. A closer look at the literature revealed no trend in the growth of deployment of either typology of tasks.

To better capture task relevance within crowdsourced studies, we have performed a refinement of the elementary/synoptic categorization using Munzner's $\{action, target\}$ task typology [Mun14]. While elementary tasks are easily expressed as $\{action, target\}$ pairs, synoptic tasks can only be expressed through lists of *({action, targets})*. Synoptic tasks, by definition, capture more complex actions such as: decision making, learning, prediction, and comprehension of a pattern & trend. These entail sequences involving more than one action, between *analyze, query* and *search*, to happen concurrently [DKSN11, YLZ14, FDPH17, KRH17]. The refinement highlighted several interesting aspects: i) a dominance of *search* and *query* actions and of *all data*, *attributes*, *network data* as targets [JRHT14, OJ15, BNK*16, KDMW16, VC17], ii) no study focusing on *spatial data* as targets. Furthermore it was noticeable how several studies split synoptic tasks into elementary components such as *analyze*, *search* and *query* with each component being analyzed within either a focused task or a separate experiment [HB10, HAS11, IFBB12, MPOW17]. This preference towards elementary tasks, might be explained by the necessity to create microtasks.

Gadiraju et al. [GKD14] found *time required* as one of the three criteria used by crowdworkers to select tasks (the other two being *interestingness* and *monetary incentives*). The use of mixed type of tasks, however, emerged later in the timeline possibly highlighting a maturity in the use of crowdsourcing platforms that resulted in attempts at developing more complex testing scenarios. Task complexity and its effect on crowdworkers performance and behavior is, however, still an open question.

### 4.3. Participants

One of the main motivations for running crowdsourcing experiments is the larger number and wider variety of participants that can be recruited. A wider variety of participants implies both access to people more representative to humanity and the global population as well as more peculiar cases that require special care. Below, we report information on study participants that were collected in our review; participant numbers, valid responses, age, gender, culture, abilities, and pre-tests, as well as measures on worker prerequisites.

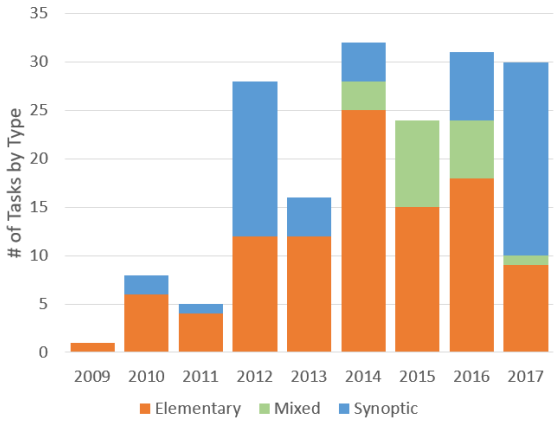


**Figure 4:** *Task distribution across year by type: Elementary, Synoptic, and Mixed.*

**Total Number of Participants:** Participant numbers in the experiments we analyzed ranged from 17 to 1687 [HYFC14] with an average of 199 participants per study and a median of 115 (Figure 5). 18 studies did not report any participant numbers at all.

**Numbers of Participants per Condition:** Larger participant numbers facilitate between-subject designs where each participant is only taking part in one condition (Section 4.1); participants have lower workloads and experiments can have more conditions. Colors in Figure 5 indicate the type of study design (within, between, mixed). Not unsurprisingly studies with a between-subject design include larger numbers of participants. The average reported number of participants per condition was 53. On average, these studies took around 16 minutes (time reported in only 73 experiments).

**Valid Responses:** Some studies removed results from participants that they found flawed. For example, results were removed where participants were too fast or too slow because this might indicate that workers were gaming the study or working on other tasks in parallel. Only a slight majority of studies (53%) reported on measures to check for invalid responses by, e.g., setting valid minimum and maximum times, excluding participants through catch-questions or through a minimum performance level. On average, these studies reported 92% valid participant responses.

**Gender:** Larger participant numbers can lead to more homogeneous samples for some characteristics, but can introduce diversity with respect to other characteristics. For example, in the 37.6% experiments that reported on participants' gender, an average 47.63% were female. Distribution of gender varied with a few as low as 29% and others as high as 77.5%. None of the studies reported on any other gender than 'male' or 'female.'

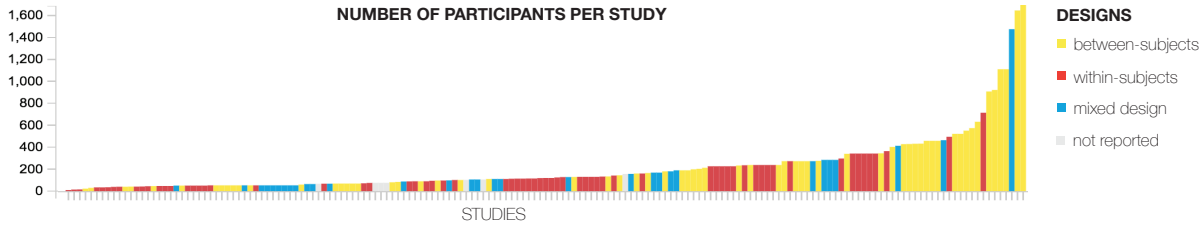


**Figure 5:** *Distribution of participant numbers (vertical) for all studies (horizontal). min = 17, max = 1687, mean = 199, median = 115.*

**Age:** Only 35% of the studies reported on workers' age. Ages ranged from 18 (a necessary condition to participate on some platforms, e.g., AMT) up to 67. Nine studies reported ages as high as 80 years old. Average age mean in the studies was 32.7 years old (min = 23.0, max = 37.7).

**Culture:** Lab experiments mostly rely on the local population, which is usually more homogeneous with respect to culture and language. In crowdsourcing studies, diversity can be introduced by the cultural origin of participants. Today, access to AMT is available to workers from all countries as well as to experimenters from 43 countries[§]. Studies from 2014 and earlier reported that at that time, most workers where concentrated in the U.S. [Ipe10a, Ipe10b] but a significant worker population exists in India, which can have implications on tasks [GMHO14]. For example, studies can involve people with a lower literacy rate and lower proficiency in English. A lack of proficiency in English can lead to difficulties in understanding tasks, experiment instructions, and predefined answers, as well as hinder expression when required to enter free text. Only 71 experiments (37 %) reported on the origin of workers. Out of these experiments 66 had participants from only the United States, and 5 indicated inclusion of workers from other countries. The remaining 63% of experiments did not report on any participant origin.

Of equal importance to experiments in visualization are biases with respect to general educational background, visualization literacy, and cultural codes. To track for diversity, crowdsourcing experiments have employed tests and specific questionnaires as follows.

**Colorblindness:** Only 31 experiments (16%) reported to have conducted a colorblindness test to prevent colorblind participants from participating. However, 14 experiments relied on self-indicated answers for color-blindness; 17 studies used colorblindness tests (e.g., [HRR45] used in [CAFG12]). Turton et al. [TWS*] address the issue of potential contamination of study results from color impaired participants.

**Pre-Test:** For proficiency in English, four experiments (all in one paper [BEDF16]) asked participants to run an intermediate English reading comprehension test. More specifically, one experiment [TTvE14] required workers to obtain a numeracy scale of higher of equal to 4.4 in numeracy test developed by Fagerlin et al. [FZFU*07].

[§] https://blog.mturk.com/mturk-is-now-available-to-requesters-from-43-countries-77d16e6a164e, last accessed Apr 15, 2018

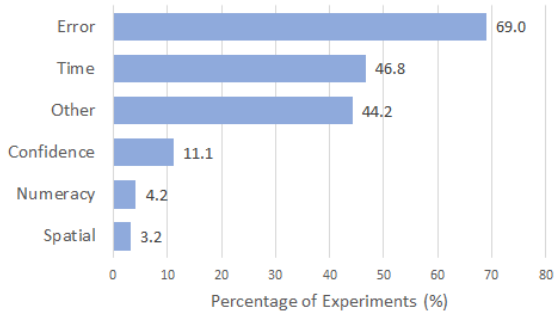


**Figure 6:** *The percentage of experiments that used each one of the mentioned measures and metrics as dependent variables.*

**Prerequisites:** Platforms such as AMT and CrowdFlower provide general tracked measures on their workers' performance. For example, 82 experiments (43%) reported to have selected workers according to such measures. The most popular criterion (34 experiments, 20.1%) was a HIT approval rate between 95%-99% and sometimes more than 10,000 HITs approved (for AMT; 5 experiments) or level 3 (for Crowdflower; 4 experiments). Some experiments (18) required participants to be based or live in the USA, another set of experiments (7) all in the same paper [BKP15] required their workers to live within an English speaking country.

### 4.4. Measures and Metrics

Figure 6 shows the popularity of different measures and metrics among the surveyed experiments.

**Error, Time, and Confidence**: As in most laboratory user experiments, error was the most popular dependent variable in the surveyed user experiments followed by time (error: 133, 70%; time: 89, 47%). Yet, time is typically considered an unreliable measure for crowdsourcing experiments [BLB*17, HB10]: differences in the workers' hardware or software could lead to different latency times; workers could take unexpected breaks or start multiple online tasks simultaneously. Remote logging of worker's actions have been proposed (e.g., [BKH14, KBB*15]) and could be used to verify whether the collected time is reliable, but analyzing such logs for a large sample population is challenging. The error and time measured for an experiment in the laboratory could also be different from those measured for the same experiment but using crowd-

sourcing [MDF12]. This may be due to factors such as crowdworkers typically having a more diverse background than participants of laboratory experiments (see Section 4.3) and different motivational drivers, such as to complete highly paid tasks with minimal cognitive effort (see Section 4.5). A small number of experiments (23, 12%, e.g., [MGGF16, MDF12, CAG13]) also measured the participants' confidence and analyzed it with respect to their error.

**Individual Abilities**: The use of psychology tests to measure and assess the effect of individual abilities, such as numeracy and spatial abilities, on user performance is becoming more popular. Experiments typically used such tests to measure numeracy (8, 5%, e.g., for reasoning about visual statistical data [OPH*16]) or spatial abilities (9, 5%; e.g., for visualizations encoding information by area [MDF12]). One experiment measured graphical literacy to assess the effectiveness of visual representations with respect to textual ones [DBH14].

**Other**: Seventy five (40%) experiments collected other less popular dependent variables, such as user preference (11, 6%, e.g., [ZRH12, HDR*13, IFBB12]), visualization quality [KDMW16, YLZ14], engagement and enjoyment [KL16, MHRL*17, HRC15], as well as textual annotations and perceived insights [YLZ14, FDPH17, WGS*13]. The type of dependent variables also depends on the aims of the experiment. For example, memorability was measured to assess the recall of different types of visualizations (e.g., [BVB*13]) and interaction logs were used to assess how a visualization was used [DKSN11, KBB*15, BKH14, SBJS15, VC17, FDPH17]. In other cases, perceptual measures, such as the just noticeable difference, were measured to model how accurately differences in visual stimuli are perceived [BDM*17]. No variables were measured in cases when the task was for instance to draw or lay out a diagram (e.g., [vHR08, KDMW16]).

### 4.5. Quality Assurance

In crowdsourcing experiments, experimenters need to control for undesired participant behavior [KSW15, CIT16]. A common technique is to allow participation only if the worker has an acceptable performance profile on the crowdsourcing platform (Section 4.3). The environment of crowdsourcing experiments is not as controlled as in laboratory experiments, so sample populations of crowdsourcing experiments are large to outweigh noise in the collected data (Section 4.1). Some test questions (dummy questions) are also not included in the data analysis to reduce possible learning effects (Section 4.1).

**Inattentive Participants**: The inclusion of consistency and attention checks in the experiment is important but not always reported (reported by only 11% of the papers). At times, two methods (e.g., [HSF*13, CAG13]), or three or even more (e.g., [CAB*16, MPOW17]) were used to identify inattentive participants (Figure 7). Non-InfoVis communities suggest the use of "gold standard" questions (e.g., 'Did you get a heart attack while completing this task?') [BOZ*14]. However, crowdworkers could easily get used to these standard questions. In fact, none of the surveyed experiments reported to have used such questions. Instead, experiment-specific catch questions, which participants should know how to answer correctly if they completed the task attentively, are more commonly being added to the experiment (17%, e.g., [BKH14, ACG14, PRS*15, CH17]).

In other cases, data from participants who answered too quickly compared to other participants or a fixed time threshold are excluded from analysis (3%, e.g., [HP17, PVF14]). Similarly, other experiments have excluded data from participants who: completed a question under three seconds more than three times [KL16]; received a negative score [BEDF16]; provided less than 150 characters for the requested textual description [KBB*17]. For some experiments, it was possible to automatically identify arbitrary answers [ZRH12] or track mouse behaviour [CAG13], and filter out invalid data accordingly.

Other experiments used instructional manipulation checks [OMD09]: a psychology technique whereby the text of the question instructs the participant to do something different from the affordances that are apparent. For instance, the text could ask the user to click somewhere else (e.g., on a specific region of the visualization) and not the provided response buttons (e.g., [RSC15]); participants that click on the buttons are considered inattentive and their data is excluded from analysis. One paper presented its experiments as micro-games [MPOW17]: participants were initially informed that a reward will be granted only if a minimum of 30% of the questions were answered correctly, and then shown the score halfway through the experiment.

**Multiple Participation**: Besides detecting inattentive participants, participants who look into or partially try the experiment while exiting without completing it, should not be allowed to re-take the experiment any time later, otherwise noise is introduced. This should be controlled by the experimenter as currently crowdsourcing platforms, such as AMT, do not provide such functionality. Yet, only

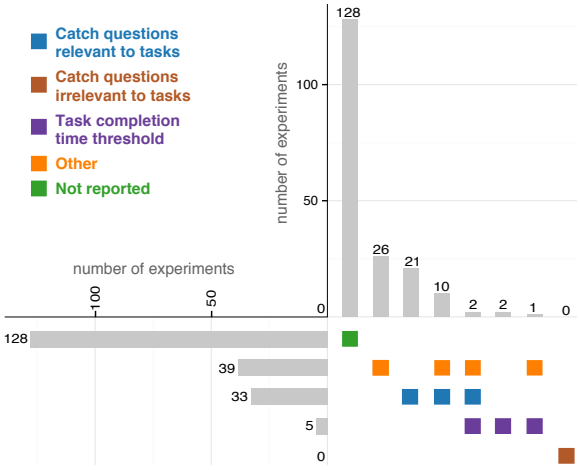


**Figure 7:** *Quality assurance techniques used by the surveyed experiments to identify inattentive participants. Horizontal bars: number of experiments that used the mentioned technique (but not necessarily exclusively). Vertical bars: number of experiments that used a specific combination of techniques.*

32% of the experiments we reviewed, reported whether this was controlled for or not.

**Payment and Bonus**: It is commonly good practice to keep crowdsourcing experiments as short as possible (on average, 16 mins; see Section 4.1) and to pay the participants adequately based on the experiment's duration. Studies show an effect of incentive on quality and measured variables (e.g., [SM13, HSH*14, HB10]), and that the greater the duration of the task, the higher the expected pay [HB10]. Yet, other studies argue that quality is better achieved through intrinsic rather than extrinsic motivators (e.g., [RG15]). The majority of the experiments opted for an extrinsic motivator: 173, 72% provided a monetary reward for the experiment or each trial; 13, 7% provided a bonus; 3, 2% had a raffle instead of a monetary reward. Interestingly, 48 experiments (25%) did not report whether the workers got any renumeration for their work, despite the various ethical principles that should be considered when designing incentives [MCG*17, SHC11].

### 4.6. Reproducibility

Reproducibility is an important aspect of any evaluation. Several crowdsourcing studies have demonstrated that crowdsourcing is successfully able to reproduce previous lab studies [HB10, OJ15], thus validating it as an approach. Having access to the original experiment materials and (anonymized) raw experiment data greatly facilitates reproducibility.

Access to the raw experiment data allows for a more detailed comparison of where results may diverge, as well as validation of analysis techniques. For instance, using the empirical data collected from a series of InfoVis crowdsourcing experiments [HYFC14], Kay and Heer [KH16] defined a model of humans' perceptual precision of estimating correlation of different visualizations, without conducting the empirical experiments themselves. Similarly, a new worker interaction loggers [BKH14] was evaluated on a previously conducted InfoVis crowdsourcing experiment [MDF12] and the data collected by the two experiments were then compared to determine the effectiveness of the logger.

In the case of lab based experiments access to experiment material may be difficult due to hardware requirements, or experiment software being developed for a specific platform or with specific dependencies. The web based nature of crowdsourcing experiments means that it is easier to make experiments materials available.

Unfortunately, only 17 (21%) of the surveyed 82 papers made their empirically collected data publicly available, and only 19 (23%) shared their experimental material. It is currently unclear what type of resources should be shared and how they should be shared. For instance, while some papers shared all their experiment material and collected data (e.g., [BDM*17]) including the scripts used for their data analysis (e.g., [FDPH17]), others shared only the stimuli and the questions (e.g., [PQMCR17]) or only the code of their online experiment (e.g., [SOK*16]). Such experiment resources are currently shared in different ways, for instance: as the paper's supplementary material (e.g., [KBB*17, MPOW17]); in a GitHub repository (e.g., [KS16, LFK*13, FDPH17]) or on a project webpage (e.g., [HRC15, BDM*17]) whose web address is provided in the paper; on a project webpage whose web address is not provided in the paper but advertised in other ways such as social media or during the presentation of the paper at the conference [MDF12]. When these resources are provided on a webpage, the authors should ensure that its accessibility is maintained. For instance, the web address provided in a small number of the surveyed papers was not functional. It might thus be sensible for the community to have a common shared repository, ideally using a version control system like GitHub, where all experiment data and material can safely be stored, shared, reused or even improved by other researchers.

## 5. Reflections on the Key Findings

Our goal of this review was not to prescribe any specific methodology for designing and execution crowdsourcing experiments but to understand how the visualization community has been using and reporting crowdsourcing for evaluation purposes. In this section, we reflect on the main findings from the survey analysis.

### 5.1. Reporting Issues

As described in Section 4, our analysis revealed that many papers failed to report important details about the experiment. In some cases, missing details not only make the study not reproducible but also make it difficult to assess whether the study was properly designed and executed, and whether the derived findings are reliable. For example, some papers did not specify the basic study design (e.g., whether it was a within-subjects or between-subjects design). Lack of reporting is spread across the entire timeline and data do not allow for any conclusion to be drawn with respect to better (or worse) experimental practices.

A checklist is commonly used to reduce failure or errors by potential limits of human memory and attention [GL10]. To ensure consistency and completeness, we provide a checklist researchers can use when reporting crowdsourcing experiments (Table 4). Furthermore, reviewers who need to review papers reporting on crowdsourcing experiments can use this checklist to ensure that the authors properly report necessary details. We, however, note that this checklist is not exhaustive in that it covers only core elements. The highlighted elements are primarily based on the six aspects that Borgo et al. identified as critical for the success of a crowdsourcing experiment for visualization research (see Section 2.5 and Table 1). For every element, we then identified factors that are important for any experiment particularly the ones using crowdsourcing based on our experience and this review. We encourage researchers to report other details, including for instance, whether they conducted a pilot study, if the supplementary material is available, and any information that is relevant to the study. We also encourage them to refer to the corresponding Dagstuhl book chapters [APH17] to inform the study design process.

To tackle some of these issues, we propose a report form for crowdsourcing experiments (template with blank space shown in Table 5 and an example of its use in Table 6). Authors can use the form to summarize each experiment as a one-pager and submit it as an appendix or supplementary material of their corresponding papers. This form can: (i) act as a detailed record of each experiment; (ii) allow continuous surveying of crowdsourcing practices

| **Study Design** |
|---|
| ☐ between-, within- or mixed |
| ☐ platform (e.g., AMT, CrowdFlower, self-hosted website) |
| ☐ system requirement (e.g., minimum screen resolution) |
| **Tasks (and Datasets)** |
| ☐ task description |
| ☐ number of tasks and trials |
| ☐ dataset description |
| **Participants** |
| ☐ recruitment method (e.g., AMT, email broadcast) |
| ☐ recruitment criteria |
| ☐ study specific (e.g., English native speakers) |
| ☐ platform specific (e.g., AMT: min. # of HITs approved, HIT approval rate) |
| ☐ number of participants |
| ☐ initially recruited, dropped, & finally analyzed |
| ☐ (if between-subjects or mixed design) per condition |
| ☐ (if participants were dropped) reasons to drop participants |
| ☐ basic demographic information (e.g., gender, age, education) |
| **Study Procedure** |
| ☐ consent |
| ☐ how participants were trained |
| ☐ average length of the study session |
| **Quality Assurance** |
| ☐ the method to avoid random clickers |
| ☐ the method to prevent multiple participation |
| **Compensation** |
| ☐ monetary, voucher, or no compensation |
| ☐ (if monetary or voucher) amount |
| ☐ if and how much bonus was provided |
| **Measures and Metrics** |
| ☐ measures (e.g., time, error, preference) |
| ☐ analysis method |
| ☐ main findings |

**Table 4:** *Checklist for reporting crowdsourcing experiments*

with minimal effort in forthcoming years; (iii) assure reviewers and readers the quality of the collected data; and (iv) ensure reproducibility of the study. Similarly, but for studies in general, Haroz provides an online template[¶] describing the minimum set of information that should be reported. In addition to the details mentioned by Haroz, our form covers aspects that are important for crowdsourcing experiments (see Section 2.5 and Table 1).

[¶] http://steveharoz.com/public/experimentmethods, last accessed Apr 15, 2018

### 5.2. Paper Types

All of the papers we surveyed had at least one experiment that used crowdsourcing for information visualization evaluation. Objects of evaluation ranged from pure perceptual and cognitive evaluation of visual channel and novel visual layouts, to learning to system usability and effectiveness in supporting analytical tasks.

A large majority of the papers were indeed evaluation papers with one or more experiments. Typically, all experiments were conducted using crowdsourcing (e.g., [MPOW17, HB10, HAS11]). In some cases, a pilot experiment was conducted in a laboratory setting (e.g., [MDF12]), while others conducted the pilot study using crowdsourcing (e.g., [HVH*16, RSC15, YLZ14, LH13]). A laboratory pilot study is typically preferred for the experimenter to better observe the participants and get direct feedback from them regarding the instructions, data, tasks and visualizations. However, testing the experiment in a crowdsourcing environment could also be helpful as the experimenter can get a better idea as to whether the various important components of the online experiments, particularly those designed specifically for the crowdsourcing experiment, such as qualification tests for participation and quality assurance tests, are working as expected in an environment that is similar to the actual study from which the findings will be derived.

The majority of the evaluation papers evaluated the perception of different types of visual variables (e.g., [HB10, IFBB12, ZCY*11]) or interactions (e.g., [BEDF16]) for different types of visualizations, such as: bar charts (e.g., [SHK15, TSA14]), parallel coordinates (e.g., [ZRH12, KL16]), scatterplots (e.g., [GCNF13, MPOW17]), time visualizations (e.g., [CAFG12, ACG14]), hierarchical visualizations (e.g., [BNK*16, VC17]), graph visualizations (e.g., [KDMW16]), surface visualizations (e.g., [SSG16]), and visualizations for narrative (e.g., [HDR*13]).

A number of other papers presented either a novel visualization technique (e.g., [KBB*17, PVF14]) or an interaction technique that could be used within existing visualization (e.g., [FDPH17, BKP15]). In both types of papers, crowdsourcing was used to evaluate the new technique against a baseline. Papers that presented a novel visual analytics systems used crowdsourcing to evaluate how well the system supported the participants in their targetted analytics tasks (e.g., [ZGWZ14, OJ15]). A small number of paper focused on a specific application domain such as social media (e.g., [HAS11]) and predictions on mobile systems (e.g., [KKHM16]).

Some papers had a very different type of contribution. For instance, a paper presented a new user interaction tracking method for online experiments and used crowdsourcing to replicate a previous crowdsourcing experiment to verify whether their method provides new insights on the reasoning process of the participants [BKH14]. One paper investigated how users learn about data visualization on online framework [KL16], while another paper assessed how annotations on visualizations can help crowdworkers produce better textual explanation of the visualized information [WHA12]. Crowdsourcing was also used to first calibrate the weights of a cost function that captured the effectiveness of a scatterplot design for different data analysis tasks, and then empirically compare the optimized designs to a number of commonly used default designs [MPOW17].

**Study:**
**Experiment:**

| | |
|---|---|
| Study Design | Type ○ between-subjects ○ within-subjects ○ mixed-design, with .....<br>Independent Variables .....<br>Dependent Variables .....<br>Covariates ..... |
| Experiment Data | Source .....<br>Specifics .....<br>Variations ..... |
| Tasks | Elementary ..... Synoptic ..... Mixed .....<br>Variations ..... |
| Visualizations | Type ○ static ○ interactive ○ other .....<br>Specifics ..... |
| Test Question Format | Question asked .....<br>Response type .....<br>Associated data, task, visualization ..... |
| Participants | Expected number per independent variable ..... in total .....<br>Requisites ..... |
| Quality Measures | Attention detectors .....<br>Multiple participation .....<br>Others ..... |
| Motivators | Intrinsic .....<br>Extrinsic ..... |
| System Requirements | Hardware .....<br>Software ..... |
| Procedure | Training .....<br>Dummy questions (not in training) .....<br>Test questions number per participant ..... ordering .....<br>Quality assurance questions .....<br>Abilities tests .....<br>Questionnaires .....<br>Steps ..... |
| Data Collection | Method .....<br>Managed by ..... Time period .....<br>Testing before actual study ..... |
| Collected Data Analysis | Average experiment length ..... Number of participants (before filtering) .....<br>Excluded participants with low quality data ..... for other reasons ..... in total .....<br>Included participants per independent variable ..... in total .....<br>Included participants demographics .....<br>Statistics methods used ..... |
| Reproducibility | Experiment Material .....<br>Collected (Anonymized) Data ..... |

**Table 5:** *A template of the form that could be filled in and provided as supplementary material to adequately report all the details of an InfoVis crowdourcing experiment. Table 6 provides an example of how the details of an experiment should be recorded using this form.*

### 5.3. Study Design and Task Types

Using Andrienko and Andrienko's task classification [AA05], our analysis revealed a preference towards elementary tasks (Section 4.2). Several factors may influence the choice of task type, ranging from crowdsourcing platform limitations to control on the quality of the collected data. Critical to crowdsourcing studies is the ability to attract high volumes of individuals that meet the specific characteristics required by the experiment. Several strategies can be employed to achieve this target. Gadiraju et al. [GKD14] identified the following three factors that typically influence the crowdworkers' task selection process: *time required* to perform a task, *monetary incentives* and *interestingness*. Monetary incentives might appear as the only factor workers take in account when choosing a task. However, Deng et al.'s [DJ16] interviews with crowdworkers revealed otherwise. Once the workers ensure that the monetary incentive is adequate for the task length and complexity, they then consider other aspects such as: *simplicity*, *variety*, and *significance* for completing the task.

Significance is related to interestingness, as workers typically select tasks which they think will have a broad impact. For instance, a number of workers who took part in an experiment that assessed visualizations to help untrained people reason about Bayesian problems thanked the experimenters for providing them with information about breast cancer and the mammography's false alarm rate [MDF12]. This example explains why crowdsourcing experiments should ideally select study data that has a meaningful context and that the workers can easily relate too. If no realistic context is provided, workers might not select the task or lose trust in the presented information; trust is important for a story to be conveyed effectively [Gla09].

Workers provided two different definitions for simplicity: (i) a task with a short explanation that is easy to learn and easy to repeat multiple times; (ii) a task with standard system and hardware requirements and thus highly *accessible*. Elementary tasks meet the criteria of both simplicity, as in Deng et al. [DJ16], and short time duration, as in Gadiraju et al. [GKD14]. A complex (synoptic) task is typically made up of smaller and simpler elementary tasks, as demonstrated in Section 4.2 when Munzner's [Mun14] task taxonomy was used to refine our task types. From a data analysis point of view, simpler tasks allow for more control over effects of confounding factors.

As crowdsourcing gained popularity, crowdsourcing visualization experiments with synoptic and mixed task types became more popular (Figure 4). This trend can be due to a maturity in the acceptance of crowdsourcing use for evaluation and academic research, leading to openness to risk. Additionally, crowdsourcing platforms are now providing improved features to for example control recruitment and identify workers with a high quality task completion history. Nonetheless, task simplicity and duration still remain a hard constraint for experimenters to meet when their studies involve tasks that are more complex than the ones typically preferred by workers.

From a study design point of view, simplicity and time duration have significant effects on design choices. Simpler tasks allow for fewer constraints on equipment and platforms, thus increasing accessibility of a task to workers. Within-subjects design is often not possible in a crowdsourcing study due to the experiment length, which could be the reason why most crowdsourcing studies in our community have opted for a between-subjects design. Yet, even if a between-subjects design is chosen, the number of trials with the same combination of study factors is often constrained, as the more the trials, the more quality assurance measures need to be included in the experiment to check the worker's attention at different stages of the experiment.

Quality assurance measures together with clear instructions and proper training (due to no direct interaction between participant and experimenter), abilities tests and questionnaires (since experimenters cannot know or predict the skill set or background of the workers) could significantly increase the experiment's length, complexity and also engagement (discussed also in 'Study Design' in Table 1). Official and well-tested abilities tests (e.g., cognitive tests [EDH76] such as the test to measure spatial abilities) are typically long. Thus experimenters that need to measure such abilities have to either conduct the experiment in a laboratory setting or not measure such abilities or include only one part of the test in the crowdsourcing experiment (e.g., part 2 of the Paper Folding Test to measure spatial abilities [MDF12]). Yet all of these options could have undesired effects on the experiment especially when the results of such tests are crucial for the study objectives.

If on the other hand full lengthy tests are included in a crowdsourcing experiment, then due to the length of the experiment, fewer crowdworkers will select the task, participants will be less engaged and their performance will not meet expectations. Decrease in participants' engagement also reduces the likelihood that the worker completes the experiment, thus slowing down data collection and study completion. If experimenters cannot reduce the experiment length and still wish to use crowdsourcing, then either the payment is increased [HB10], or incentives to maintain participants' engagement and interests are devised (e.g., deploying the experiment as a game [MPOW17]), or data for repetitions of combination of study factors is collected from a larger and more diverse pool of participants. Studies that require participants with specific requisites (e.g., full color vision) could, in some cases, take advantage of the large volume of crowdworkers and increase the size of their participants pool, such that the uncertainty introduced by participants not meeting the requisites is reduced. In such cases the crowdsourcing experiment would not test for the requisites and thus, keep the experiment within a constrained length.

Unfortunately, a large number of papers did not adequately report on details, such as the number of trials and the methodology used to filter participants. Thus, we could not test for any possible correlations between design choices and studies contextual factors.

### 5.4. Limitations

As with all surveys, we could only report on experiments that matched our search criteria and our focus of this survey. Moreover, there might have been papers that were not properly archived and thus were not captured in our search. Alternatively, limitations in the surveyed papers might have led us to inaccurately report and analyze their experiments.

The possibilities and options to conduct crowdsourcing exper-

iments might be much wider than what we could report. For example, there might be other strategies to capture inattentive participants and other crowdsourcing platforms that were not used in the experiments we surveyed. Similarly, our dimensions could be extended in the future, especially if new methods, technologies, or trends emerge. The taxonomy we used to categorize the tasks adequately matched our survey goals and the survey experiments. Yet, other task taxonomies could be used for other objectives (see Section 4.2). After reviewing all the experiments, we realized that 40% of the experiments, particularly the more recent ones, used measures that go beyond the traditional error, time and confidence metrics. If this trend of novel metrics (e.g., engagement and enjoyment) keeps gaining popularity, then follow-up reviews could consider splitting up our current other measures dimension into more specific types of metrics.

The focus of our survey is on the use of crowdsourcing for visualization evaluation. Other research fields such as Human Computer Interaction, Psychology, and Social Sciences have addressed the use of crowdsourcing within their respective remit. A comparison of our findings against these body of work is beyond the scope of this survey. We hope that our results will be a useful tool for future works attempting to perform such comparison.

## 6. Opportunities and Challenges

In this section, we discuss several opportunities for improvements and challenges in using crowdsourcing for visualization research which we drew from the analysis of 190 experiments.

### 6.1. Find the Right Workforce

The demographics and background of crowdworkers is typically very diverse. Thus, it could be easier for experimenters to find participants with the required the skill set. For instance, experiments might evaluate the effectiveness of visualizations that are specifically designed for participants that are color-blind or have a specific cultural background or domain knowledge (e.g., nurses, designers, managers). In other cases, experiments might be conducted to evaluate the effect of age or individual traits on user performance in completing a specific visual analysis task. As long as the adequate checks are carried out to ensure all participants qualify to the experiment's requirements, then crowdsourcing can open up great opportunities to visualization researchers.

To attract crowdworkers that meet the experiment requirements, a number of papers have reported on different strategies of how experimenters should advertise their experiments and how crowdworkers select the tasks they want to complete. For instance, workers typically select tasks that are well-paid and recently posted on the crowdsourcing platform [CHMA10]. Thus, providing competitive pay and posting tasks in small batches is typically considered good practice [BKMB12]. It is also common for workers to complete a chain of tasks in one session. This can lead to cognitive overload and an impairment in the worker's performance [CIT16]. So attention checks should always be included in the experiment, and the experiment should ideally be advertized at a time when other recently posted tasks require a worker skill set that is different from that of the experiment.

Workers spend more than 25% of their time searching for tasks that they want to complete [KDM*14]. Thus, in the future, it would be helpful for crowdsourcing platforms to either order the tasks based on their cognitive demands [CIT16] or automatically fetch and suggest tasks to the worker [LRMB15, CIT16, GCB16] based on a self-assessment report filled in by the worker [GFK*17]. If such a self-assessment report also includes the results of the ability tests (e.g., colorblindness, visual literacy) that the worker has completed for previous tasks, and these reports are accessible to all experimenters, then experiments could be shorter, more focused on the assessed task, and less cumbersome for the worker. Adopting such a mechanism for the experiments in our community would not only help us find participants, but also ensure that our participants fit our requisites. Unfortunately, current popular crowdsourcing platform do not support such profiling of worker abilities.

Another possibility is to take to inverse approach, and integrate crowdsourcing functionality into an existing online platform, which has the target skill set or intrinsic motivation desired by the experimenters. For example, van Ham and Rogowitz integrated crowdsourcing functionality into the into the (now defunct) "Many Eyes" collaborative visualization website [vHR08], for their study of perceptual organization of network layout. Using a visualization website in this way meant that it was more likely that the participants would be motivated to perform the task correctly and would have, at a minimum, a basic understanding of networks.

### 6.2. Consider Other Possible Covariates

Studies from the outside of the visualization community have shown that the following factors could affect the error collected from crowdsourcing experiments: interface design, usability and simplicity [KRDT10, FKTC13, KRDT10, FKTC13, RD14, BAC14]; mobile usage for conducting experiment [FZFM17, IH17]; age [KKMF11]; incentives [SM13]; narrative [DBD17]. Yet, none of the 190 experiments we surveyed took these factors into account as possible covariates, which either confound or interact with the measured dependent variables.

### 6.3. Adopt Novel Quality Assurance Techniques

A number of novel quality assurance techniques that have been adopted in other domains could be considered for InfoVis research. For instance, the crowdsourcing experiments could be devised in the form of a game to increase engagement and the likelihood of collecting data from attentive workers (e.g., [CK13, KBB*15, LSKB13]). Workers' answers could be compared with those provided by the rest of the sample population [SL13] or peer reviewed [ZDKK14, WYF*17]. Low quality answers could be filtered out based on the behavior of trusted workers [KZ16] or the predicted answer to a question [GG17] or task completion time [WFI11, KDTM16]. Instead of crowdsourcing, "friendsourcing" could be used to outsource experiments to a group of reliable individuals that are socially connected and are known to complete experiments accurately possible due to intrinsic rather than extrinsic motivators [BTS*10].

Workers' interactions have been logged in InfoVis experiments to assess how a visualization was used (Section 4.4), but not

used to track inattentive workers. Yet, studies indicate that malicious behavior is traceable from logs of example mouse clicks and movements, key presses, scrolls, and changes of focus [GKDD15, RK11]. Manually analyzing the interaction logs of a large group of participants is challenging, but tools are now available to visualize such data [HGM11, WZT*16] and analyze it [RK12].

### 6.4. Leverage Intrinsic Motivators

An HCI study demonstrates that crowdworkers are still eager to complete uncompensated online experiments [RG15]. The main intrinsic motivators were to compete with others, to learn about themselves, to improve specific skills, or to be fascinated by the work. However, most experiments in our survey did not consider intrinsic motivators. One paper [MPOW17] in our survey attempted to intrinsically motivate the workers, by showing their score halfway through the experiment, but yet a monetary reward was provided if at least 30% of the questions were answered correctly. More time and effort is required to devise an online experiment with effective intrinsic motivators (e.g., an experiment in the form of a game). User engagements is still an open research questions. However, if as a community we could systematically share code of our online experiments and make it publicly available, then other researchers could build on top of or integrate parts of the code in their own online experiments and save a considerable amount of time and effort in devising, implementing and testing such intrinsic motivators. This strategy has been adopted widely and successfully in the Psychology community, of which Psychtoolbox [Bra97] is an example.

### 6.5. Share Resources with the Community

Implementing and testing an online experiment that works as expected on different browsers and machines requires time and effort, particularly when devising novel motivators, quality assurance techniques, measures and tests to check the workers' hardware, software and requisites. So sharing our experiment code, so that other researchers can use it as a template for their experiment, will help our community conduct higher quality experiments more effectively and efficiently. This will also ensure reproducibility (Section 4.6) and facilitate future empirical comparisons [KLB*16]. For instance, one paper [BKH14] evaluated a novel worker interaction logging mechanism by using the same publicly available code as a previous crowdsourcing InfoVis experiment [MDF12]. Since the collected data of the previous experiment were also publicly available, the results of the two experiments could be compared and the effect of the new technique assessed.

Having a system, where all the resources in connection to a paper could be stored, could facilitate sharing and reproducibility, as is becoming more popular in other fields like bioinformatics with systems like MicroReact[‖] and Open Science Framework[**] and other recommended repositories (e.g., by PLOS ONE[††]). A workforce with specific visualization requirements could possibly be shared in a similar manner or as discussed in Section 6.1. In doing so, various ethical principles need to be taken into account [MCG*17].

### 6.6. Leverage Advances in Technology

Popular crowdsourcing platforms like AMT are missing important features to, for instance, avoid multiple participation, balance gender between experimental groups, and check participants hardware and software requirements. Thus, such features have to be handled by each experiment. Sharing our experiment code with the community could encourage code reuse (Section 6.5). Another option is to use new emerging crowdsourcing platforms (e.g., [TBRA17, CMZF17]; reviews [HJR*17, YKL11]). Back in 2008, a paper proposed a new platform to support InfoVis crowdsourcing experiments [vHR08], but this platform has not been used extensively and maybe it is time to consider developing a new platform that better supports the current InfoVis research trends, as already done for scientific visualizations [EKR16].

In addition, other communities are now using technologies, such as eye tracking [LMS*15], virtual reality [VAJO13], and emotion detectors [MKP12] (more [HJR*17]), for crowdsourcing experiments. Such technologies could also be helpful for InfoVis research and thus be explored in the future.

### 6.7. Crowdsourcing for Purposes Other Than Evaluation

A small number of InfoVis papers have used crowdsourcing to collect data, such as user generated layouts (e.g., [KDMW16, KWD14, vHR08]), providing useful insights into layout characteristics that the general population is familiar with or finds helpful. Yuan et al. [YCHZ12] use crowdsourcing to provide multiple input layouts of sub-graphs which are then merged together to generate a layout of the full graph. This is also an interesting example of crowdsourcing for collaborative visualization. The basic task in this instance concerns layout. However crowdsourcing for collaborative visualization opens up a new range of possible tasks.

Heer et al. [HVW07] describe a collaborative visualization system that allows for social data analysis of United States census data. While their work pre-dates the rise in popularity of crowdsourcing within the visualization community and they never explicitly use the term, it is a clear example of crowdsourced data visualization exploration. The tool allows user comments, bookmarks of specific views, and annotations to be shared as users explore the data. While there are no specific task provided for users in the associated user study, the act of commentating, annotating, and sharing findings of interest can be considered social exploration tasks.

Crowdsourcing can also be used to collaboratively validate data quality within a visualization system. In the Digital Humanities domain, Histograph [NMM*14] was specifically devised to allow collaborative analysis of networks created from historical data. It uses a voting system to enable discussion about the uncertain or contradictory information that is common in the application domain. In this way, the task of data validation and verification is crowdsourced to the platform users and the network is then updated based on the wisdom of the crowd.

‖ https://microreact.org, last accessed Apri 15, 2018
** https://osf.io, last accessed Apr 15, 2018
†† http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories, last accessed Apr 15, 2018

However, despite the potential range of other applications the most popular use of crowdsourcing in our community is for evaluation of static visualization, with AMT being the most popular platform (Section 4.1) even though AMT does not have all the necessary features to support our experiments, such as balancing participants' demographics among experiment groups (Section 6.6). Other communities are already using crowdsourcing for diverse purposes other than evaluation [GGMN17, HJR*17] that could be related to InfoVis: e.g., simulating human perception [RELS14], crowdsourcing clustering of graphics [SRdSR*15], collecting data for predicting website aesthetics [RYM*13], or for recommending graphics and visualizations [MVTS15].

## 7. Conclusion

We have presented the first systematic review of the use of crowdsourcing in the last decade for the evaluation of information visualization. This review brings to light a number of common practices, issues, challenges and opportunities with respect to different aspects of a crowdsourcing user experiment. Visualization researchers new to crowdsourcing will find this review instrumental in effectively designing, executing and reporting their crowdsourcing empirical user experiments. Other researchers experienced in using crowdsourcing are encouraged to reflect on our findings and implement the necessary changes in their current evaluation methodology, to ensure the correctness of their study design and execution, the quality of the collected data, and the reproducibility of their research.

We have presented a taxonomy of practices adopted at different stages of a crowdsourcing user experiment, and a checklist to support information visualization researchers in reporting important details for understanding and reproducing the experiment procedure and its outcomes. During the execution of this review, the need for such a checklist was immediately apparent as a large majority of the surveyed papers failed to report important details of the executed experiments. We purposely classified missing information as not reported, as we are aware that at times such information is not reported but still taken into account due to typical space restrictions in the main manuscript. Nonetheless, we should reflect whether such a common practice in our community is acceptable if we want to encourage the publication of high quality and reproducible research. We hope our review will help our research community to increase its consciousness in the use of crowdsourcing as a tool and have a positive impact in the development of more rigorous empirical research.

## 8. Acknowledgments

This work was initiated at Dagstuhl Seminar 15481 on 'Evaluation in the Crowd: Crowdsourcing and Human-Centred Experiments'. We would like to thank the seminar organizers, Daniel Archambault, Tobias Hoßfeld, and Helen C. Purchase, as well as all of the participants for the interesting discussions that motivated this work. We also thank the staff at Schloss Dagstuhl for hosting this seminar. We also would like to extend our special thanks to Sara Fabrikant for her initial comments and to the reviewers for their suggestions for improvement.

## References

[AA05] ANDRIENKO N., ANDRIENKO G.: *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. 7, 14

[ACG14] ALBERS D., CORRELL M., GLEICHER M.: Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2014), CHI '14, ACM, pp. 551–560. 8, 10, 12

[AJB16] ADNAN M., JUST M., BAILLIE L.: Investigating time series visualisations to improve the user experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 5444–5455. 5

[APH17] ARCHAMBAULT D., PURCHASE H., HOSSFELD T.: *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22 – 27, 2015, Revised Contributions*. Lecture Notes in Computer Science. Springer International Publishing, 2017. 1, 11

[AZM12] AHMED N., ZHENG Z., MUELLER K.: Human computation in visualization: Using purpose driven games for robust evaluation of visualization algorithms. *IEEE transactions on visualization and computer graphics 18*, 12 (2012), 2104–2113. 7

[BAC14] BOZZON A., AROYO L., CREMONESI P.: First international workshop on user interfaces for crowdsourcing and human computation. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces* (2014), ACM, pp. 398–400. 15

[BCER14] BADAM S. K., CHANDRASEGARAN S., ELMQVIST N., RAMANI K.: Tracing and sketching performance using blunt-tipped styli on direct-touch tablets. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces* (2014), ACM, pp. 193–200. 4

[BCLT14] BREHMER M., CARPENDALE S., LEE B., TORY M.: Predesign empiricism for information visualization: scenarios, methods, and challenges. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization* (2014), ACM, pp. 147–151. 2

[BDF15] BOY J., DETIENNE F., FEKETE J.-D.: Storytelling in information visualizations: Does it engage users to explore data? In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, pp. 1449–1458. 7

[BDM*17] BEECHAM R., DYKES J., MEULEMANS W., SLINGSBY A., TURKAY C., WOOD J.: Map lineups: effects of spatial structure on graphical inference. *IEEE transactions on visualization and computer graphics 23*, 1 (2017), 391–400. 10, 11

[BEDF16] BOY J., EVEILLARD L., DETIENNE F., FEKETE J.-D.: Suggested interactivity: Seeking perceived affordances for information visualization. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 639–648. 7, 8, 9, 10, 12

[BKH14] BRESLAV S., KHAN A., HORNBÆK K.: Mimic: visual analytics of online micro-interactions. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces* (2014), ACM, pp. 245–252. 7, 8, 9, 10, 11, 12, 16

[BKMB12] BERNSTEIN M. S., KARGER D. R., MILLER R. C., BRANDT J.: Analytic methods for optimizing realtime crowdsourcing. *arXiv preprint arXiv:1204.2995* (2012). 15

[BKP15] BERTINI E., KENNEDY J., PUPPO E.: Interaction with uncertainty in visualisations. 9, 12

[BLB*17] BORGO R., LEE B., BACH B., FABRIKANT S., JIANU R., KERREN A., KOBOUROV S., MCGEE F., MICALLEF L., VON LANDESBERGER T., BALLWEG K., DIEHL S., SIMONETTO P., ZHOU M.: Crowdsourcing for information visualization: Promises and pitfalls. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments* (Cham, 2017), Archambault D., Purchase H., Hoßfeld T., (Eds.), Springer International Publishing, pp. 96–138. 1, 3, 4, 9

[BNK*16] BACHER I., NAMEE B. M., KELLEHER J. D., BERTINI E., ELMQVIST N., WISCHGOLL T.: Using icicle trees to encode the hierarchical structure of source code. *EuroVis 2016-Short Papers* (2016). 7, 8, 12

[BOZ*14] BROWN E., OTTLEY A., ZHAO H., LIN Q., SOUVENIR R., ENDERT A., CHANG R.: Finding waldo: Learning about users from their interactions. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (Dec 2014), 1663–1672. 10

[Bra97] BRAINARD D. H.: The Psychophysics Toolbox. *Spatial Vision 10*, 4 (1997), 433–436. 16

[BTS*10] BERNSTEIN M. S., TAN D., SMITH G., CZERWINSKI M., HORVITZ E.: Personalization via friendsourcing. *ACM Transactions on Computer-Human Interaction (TOCHI) 17*, 2 (2010), 6. 15

[BVB*13] BORKIN M. A., VO A. A., BYLINSKII Z., ISOLA P., SUNKAVALLI S., OLIVA A., PFISTER H.: What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2306–2315. 10

[CAB*16] CHUNG D. H., ARCHAMBAULT D., BORGO R., EDWARDS D. J., LARAMEE R. S., CHEN M.: How ordered is it? on the perceptual orderability of visual channels. *EuroVis 2016* (2016). 7, 10

[CAFG12] CORRELL M., ALBERS D., FRANCONERI S., GLEICHER M.: Comparing averages in time series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 1095–1104. 9, 12

[CAG13] CORRELL M. A., ALEXANDER E. C., GLEICHER M.: Quantity estimation in visualizations of tagged text. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2013), ACM, pp. 2697–2706. 10

[Car08] CARPENDALE S.: Evaluating information visualizations. In *Information visualization*. Springer, 2008, pp. 19–45. 2

[CH17] CORRELL M., HEER J.: Regression by eye: Estimating trends in bivariate visualizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 1387–1396. 10

[CHMA10] CHILTON L. B., HORTON J. J., MILLER R. C., AZENKOT S.: Task search in a human computation market. In *Proceedings of the ACM SIGKDD workshop on human computation* (2010), ACM, pp. 1–9. 15

[CIT16] CAI C. J., IQBAL S. T., TEEVAN J.: Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 3143–3154. 10, 15

[CK13] CHANDLER D., KAPELNER A.: Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization 90* (2013), 123–133. 15

[CMFH12] COOK D., MAJUMDER M., FOLLETT L., HOFMANN H.: Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization And Computer Graphics 18* (2012), 2441–2448. 8

[CMZF17] CHEN C., MENG X., ZHAO S., FJELD M.: Retool: Interactive microtask and workflow design through demonstration. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 3551–3556. 16

[CTIB15] CHENG J., TEEVAN J., IQBAL S. T., BERNSTEIN M. S.: Break it down: A comparison of macro- and microtasks. ACM Conference on Human Factors for Computing Systems (CHI). 7

[DBD17] DIMARA E., BEZERIANOS A., DRAGICEVIC P.: Narratives in crowdsourced evaluation of visualizations: A double-edged sword? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 5475–5484. 15

[DBH14] DEMIRALP C., BERNSTEIN M., HEER J.: Learning perceptual kernels for visualization design. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (Dec 2014), 1933–1942. 8, 10

[DCS*17] DROUHARD M., CHEN N.-C., SUH J., KOCIELNIK R., PENA-ARAYA V., CEN K., ZHENG X., ARAGON C. R.: Aeonium: Visual analytics to support collaborative qualitative coding. In *Pacific Visualization Symposium (PacificVis), 2017 IEEE* (2017), IEEE, pp. 220–229. 4

[DDW11] DILLINGHAM I., DYKES J., WOOD J.: Visual analytical approaches to evaluating uncertainty and bias in crowd sourced crisis information. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 273–274. 4

[DH08] DALSGAARD P., HANSEN L. K.: Performing perceptionâĂŤstaging aesthetics of interaction. *ACM Transactions on Computer-Human Interaction (TOCHI) 15*, 3 (2008), 13. 4

[DJ16] DENG X., JOSHI K.: Why individuals participate in microtask crowdsourcing work environment: Revealing crowdworkers' perceptions. *Journal of the Association for Information Systems 17* (2016), 648–673. 14

[DKSN11] DIAKOPOULOS N., KIVRAN-SWAINE F., NAAMAN M.: Playable data: characterizing the design space of game-y infographics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2011), ACM, pp. 1717–1726. 8, 10

[EDH76] EKSTROM R. B., DERMEN D., HARMAN H. H.: *Manual for kit of factor-referenced cognitive tests*, vol. 102. Educational Testing Service Princeton, NJ, 1976. 14

[EKR16] ENGLUND R., KOTTRAVEL S., ROPINSKI T.: A crowdsourcing system for integrated and reproducible evaluation in scientific visualization. In *Pacific Visualization Symposium (PacificVis), 2016 IEEE* (2016), IEEE, pp. 40–47. 4, 16

[FDPH17] FENG M., DENG C., PECK E. M., HARRISON L.: Hindsight: encouraging exploration through direct encoding of personal interaction history. *IEEE transactions on visualization and computer graphics 23*, 1 (2017), 351–360. 8, 10, 11, 12

[FFB18] FELIX C., FRANCONERI S., BERTINI E.: Taking word clouds apart: An empirical investigation of the design space for keyword summaries. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan 2018), 657–666. 8

[FKTC13] FINNERTY A., KUCHERBAEV P., TRANQUILLINI S., CONVERTINO G.: Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI* (2013), ACM, p. 14. 15

[FSASK13] FIGUEROLA SALAS O., ADZIC V., SHAH A., KALVA H.: Assessing internet video quality using crowdsourcing. In *Proceedings of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia* (New York, NY, USA, 2013), CrowdMM '13, ACM, pp. 23–28. 7

[FZFM17] FINDLATER L., ZHANG J., FROEHLICH J. E., MOFFATT K.: Differences in crowdsourced vs. lab-based mobile and desktop input performance data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 6813–6824. 15

[FZFU*07] FAGERLIN A., ZIKMUND-FISHER B. J., UBEL P. A., JANKOVIC A., DERRY H. A., SMITH D. M.: Measuring numeracy without a math test: development of the subjective numeracy scale. *Medical Decision Making 27*, 5 (2007), 672–680. 9

[GCB16] GOULD S. J., COX A. L., BRUMBY D. P.: Diminished control in crowdsourcing: an investigation of crowdworker multitasking behavior. *ACM Transactions on Computer-Human Interaction (TOCHI) 23*, 3 (2016), 19. 15

[GCNF13] GLEICHER M., CORRELL M., NOTHELFER C., FRANCONERI S.: Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2316–2325. 8, 12

[GFK*17] GADIRAJU U., FETAHU B., KAWASE R., SIEHNDEL P., DIETZE S.: Using worker self-assessments for competence-based preselection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI) 24*, 4 (2017), 30. 15

[GG17] GURARI D., GRAUMAN K.: Crowdverge: Predicting if people will agree on the answer to a visual question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 3511–3522. 15

[GGMN17] GHEZZI A., GABELLONI D., MARTINI A., NATALICCHIO A.: Crowdsourcing: a review and suggestions for future research. *International Journal of Management Reviews* (2017). 17

[GKD14] GADIRAJU U., KAWASE R., DIETZE S.: A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (New York, NY, USA, 2014), HT '14, ACM, pp. 218–223. 8, 14

[GKDD15] GADIRAJU U., KAWASE R., DIETZE S., DEMARTINI G.: Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, pp. 1631–1640. 16

[GL10] GAWANDE A., LLOYD J. B.: *The checklist manifesto: How to get things right*, vol. 200. Metropolitan Books New York, 2010. 11

[Gla09] GLASSNER A.: *Interactive storytelling: Techniques for 21st century fiction*. AK Peters/CRC Press, 2009. 14

[GMHO14] GUPTA N., MARTIN D., HANRAHAN B. V., O'NEILL J.: Turk-life in india. In *Proceedings of the 18th International Conference on Supporting Group Work* (2014), ACM, pp. 1–11. 9

[GMN*17] GADIRAJU U., MÖLLER S., NÖLLENBURG M., SAUPE D., EGGER-LAMPL S., ARCHAMBAULT D., FISHER B.: Crowdsourcing versus the laboratory: Towards human-centered experiments using the crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments* (2017), Archambault D., Purchase H., Hoßfeld T., (Eds.), Springer International Publishing, pp. 6–26. 3, 6

[HAS11] HULLMAN J., ADAR E., SHAH P.: The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2011), CHI '11, ACM, pp. 1461–1470. 8, 12

[HB10] HEER J., BOSTOCK M.: Crowdsourcing graphical perception: using mechanical turk to assess visualization design. *Proceedings of the 28th SIGCHI Conference on Human Factors in Computing Systems (CHI)* (2010), 203âĂŞ212. 2, 5, 8, 9, 11, 12, 14

[HCL12] HARRISON L., CHANG R., LU A.: Exploring the impact of emotion on visual judgement. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (2012), IEEE, pp. 227–228. 7

[HDR*13] HULLMAN J., DRUCKER S., RICHE N. H., LEE B., FISHER D., ADAR E.: A deeper understanding of sequence in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec 2013), 2406–2415. 8, 10, 12

[HF13] HUANG S.-W., FU W.-T.: Enhancing reliability using peer consistency evaluation in human computation. In *Proceedings of the 2013 conference on Computer supported cooperative work* (2013), ACM, pp. 639–648. 2

[HGM11] HEYMANN P., GARCIA-MOLINA H.: Turkalytics: analytics for human computation. In *Proceedings of the 20th international conference on World wide web* (2011), ACM, pp. 477–486. 16

[HHR*14] HOSSFELD T., HIRTH M., REDI J., MAZZA F., KORSHUNOV P., NADERI B., SEUFERT M., GARDLO B., EGGER S., KEIMEL C.: Best practices and recommendations for crowdsourced qoe-lessons learned from the qualinet task force" crowdsourcing". 2

[HJR*17] HIRTH M., JACQUES J., RODGERS P., SCEKIC O., WYBROW M.: Crowdsourcing technology to support academic research. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments* (2017), Archambault D., Purchase H., Hoßfeld T., (Eds.), Springer International Publishing, pp. 70–95. 3, 16, 17

[HLS16] HEARST M. A., LASKOWSKI P., SILVA L.: Evaluating information visualization via the interplay of heuristic evaluation and question-based scoring. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 5028–5033. 5

[How06] HOWE J.: The rise of crowdsourcing. *WIRED* (June 2006). URL: http://www.wired.com/2006/06/crowds/. 3

[HP17] HUNG Y.-H., PARSONS P.: Assessing user engagement in information visualization. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2017), ACM, pp. 1708–1717. 10

[HRC15] HARRISON L., REINECKE K., CHANG R.: Infographic aesthetics: Designing for the first impression. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, pp. 1187–1190. 7, 10, 11

[HRR45] HARDY L. H., RAND G., RITTLER M. C.: Tests for the detection and analysis of color-blindness. i. the ishihara test: an evaluation. *JOSA 35*, 4 (1945), 268–275. 9

[HSF*13] HARRISON L., SKAU D., FRANCONERI S., LU A., CHANG R.: Influencing visual judgment through affective priming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), ACM, pp. 2949–2958. 10

[HSH*14] HIRTH M., SCHEURING S., HOSSFELD T., SCHWARTZ C., TRAN-GIA P.: Predicting result quality in crowdsourcing using application layer monitoring. In *Communications and Electronics (ICCE), 2014 IEEE Fifth International Conference on* (2014), IEEE, pp. 510–515. 11

[HVH*16] HEINRICH J., VUONG J., HAMMANG C., WU A., RITTENBRUCH M., HOGAN J., BRERETON M., OâĂŹDONOGHUE S.: Evaluating viewpoint entropy for ribbon representation of protein structure. In *Proceedings of the 37th Annual Conference of the European Association for Computer Graphics* (2016), The Eurographics Association and John Wiley & Sons Ltd. 12

[HVW07] HEER J., VIÉGAS F. B., WATTENBERG M.: Voyagers and voyeurs: Supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2007), CHI '07, ACM, pp. 1029–1038. 16

[HYFC14] HARRISON L., YANG F., FRANCONERI S., CHANG R.: Ranking visualizations of correlation using weber's law. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 1943–1952. 8, 11

[IFBB12] ISENBERG T., FEKETE J., BEZERIANOS A., BOUKHELIFA N.: Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Transactions on Visualization & Computer Graphics 18* (2012), 2769–2778. 8, 10, 12

[IH17] IKEDA K., HOASHI K.: Crowdsourcing go: Effect of worker situation on mobile crowdsourcing performance. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 1142–1153. 15

[Ipe10a] IPEIROTIS P. G.: Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students 17*, 2 (2010), 16–21. 9

[Ipe10b] IPEIROTIS P. G.: Demographics of mechanical turk. 9

[JRHT14] JIANU R., RUSU A., HU Y., TAGGART D.: How to display group information on node-link diagrams: An evaluation. *IEEE Transactions on Visualization and Computer Graphics 20*, 11 (Nov 2014), 1530–1541. 8

[JS10] JIN J., SZEKELY P.: Interactive querying of temporal data using a comic strip metaphor. In *2010 IEEE Symposium on Visual Analytics Science and Technology* (2010), pp. 163–170. 8

[KBB*15] KIM N. W., BYLINSKII Z., BORKIN M. A., OLIVA A., GAJOS K. Z., PFISTER H.: A crowdsourced alternative to eye-tracking for visualization understanding. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (2015), ACM, pp. 1349–1354. 7, 9, 10, 15

[KBB*17] KIM N. W., BYLINSKII Z., BORKIN M. A., GAJOS K. Z., OLIVA A., DURAND F., PFISTER H.: Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI) 24*, 5 (2017), 36. 10, 11, 12

[KDM*14] KUCHERBAEV P., DANIEL F., MARCHESE M., CASATI F., REAVEY B.: Toward effective tasks navigation in crowdsourcing. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces* (2014), ACM, pp. 401–404. 15

[KDMW16] KIEFFER S., DWYER T., MARRIOTT K., WYBROW M.: Hola: Human-like orthogonal network layout. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 349–358. 4, 8, 10, 12, 16

[KDTM16] KUCHERBAEV P., DANIEL F., TRANQUILLINI S., MARCHESE M.: Relauncher: crowdsourcing micro-tasks runtime controller. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (2016), ACM, pp. 1609–1614. 15

[KH16] KAY M., HEER J.: Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 469–478. 5, 11

[KHA10] KONG N., HEER J., AGRAWALA M.: Perceptual guidelines for creating rectangular treemaps. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 990–998. 8

[KHA16] KIM Y.-S., HULLMAN J., AGRAWALA M.: Generating personalized spatial analogies for distances and areas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 38–48. 4

[KHD12] KEIMEL C., HABIGT J., DIEPOLD K.: Challenges in crowd-based video quality assessment. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on* (July 2012), pp. 13–18. 7

[KKHM16] KAY M., KOLA T., HULLMAN J. R., MUNSON S. A.: When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 5092–5103. 7, 12

[KKL16] KIM H.-R., KANG H., LEE I.-K.: Image recoloring with valence-arousal emotion model. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 209–216. 4

[KKMF11] KAZAI G., KAMPS J., MILIC-FRAYLING N.: Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (2011), ACM, pp. 1941–1944. 15

[KKMF13] KAZAI G., KAMPS J., MILIC-FRAYLING N.: An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval 16*, 2 (2013), 138–178. 2

[KL16] KWON B. C., LEE B.: A comparative evaluation on online learning approaches using parallel coordinate visualization. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 993–997. 10, 12

[KLB*16] KIDWELL M. C., LAZAREVIĆ L. B., BARANSKI E., HARDWICKE T. E., PIECHOWSKI S., FALKENBERG L.-S., KENNETT C., SLOWIK A., SONNLEITNER C., HESS-HOLDEN C., ET AL.: Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PLoS biology 14*, 5 (2016), e1002456. 5, 16

[KLKY11] KIM S.-H., LI S., KWON B. C., YI J. S.: Investigating the efficacy of crowdsourcing on evaluating visual decision supporting system. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 55*, 1 (2011), 1090–1094. 7

[KRDT10] KHANNA S., RATAN A., DAVIS J., THIES W.: Evaluating and improving the usability of mechanical turk for low-income workers in india. In *Proceedings of the first ACM symposium on computing for development* (2010), ACM, p. 12. 15

[KRH17] KIM Y.-S., REINECKE K., HULLMAN J.: Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI '17, ACM, pp. 1375–1386. 8

[KS16] KOSARA R., SKAU D.: Judgment Error in Pie Chart Variations. In *EuroVis 2016 - Short Papers* (2016), Bertini E., Elmqvist N., Wischgoll T., (Eds.), The Eurographics Association. 8, 11

[KSW15] KLIPPEL A., SPARKS K., WALLGRÜN J.: Pitfalls and potentials of crowd science: A meta-analysis of contextual influences. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2* (2015), 325. 3, 10

[KWD14] KACHKAEV A., WOOD J., DYKES J.: Glyphs for exploring crowd-sourced subjective survey classification. *Computer Graphics Forum 33*, 3 (June 2014), 311–320. 4, 16

[KWS*14] KERNE A., WEBB A. M., SMITH S. M., LINDER R., LUPFER N., QU Y., MOELLER J., DAMARAJU S.: Using metrics of curation to evaluate information-based ideation. *ACM Transactions on Computer-Human Interaction (ToCHI) 21*, 3 (2014), 14. 4

[KZ16] KAZAI G., ZITOUNI I.: Quality management in crowdsourcing using gold judges behavior. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (2016), ACM, pp. 267–276. 15

[LBI*12] LAM H., BERTINI E., ISENBERG P., PLAISANT C., CARPENDALE S.: Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics 18*, 9 (2012), 1520–1536. 2

[LFK*13] LIN S., FORTUNA J., KULKARNI C., STONE M., HEER J.: Selecting semantically-resonant colors for data visualization. In *Computer Graphics Forum* (2013), vol. 32, Wiley Online Library, pp. 401–410. 11

[LH13] LIN S., HANRAHAN P.: Modeling how people extract color themes from images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), ACM, pp. 3101–3110. 12

[LHM*15] LEBRETON P., HUPONT I., MÄKI T., SKODRAS E., HIRTH M.: Eye tracker in the wild: Studying the delta between what is said and measured in a crowdsourcing experiment. In *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia* (New York, NY, USA, 2015), CrowdMM '15, ACM, pp. 3–8. 7

[LMS*15] LEBRETON P., MÄKI T., SKODRAS E., HUPONT I., HIRTH M.: Bridging the gap between eye tracking and crowdsourcing. In *Human Vision and Electronic Imaging XX* (2015), vol. 9394, International Society for Optics and Photonics, p. 93940W. 16

[LRMB15] LASECKI W. S., RZESZOTARSKI J. M., MARCUS A., BIGHAM J. P.: The effects of sequence and delay on crowd work. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, pp. 1375–1378. 15

[LSKB13] LASECKI W. S., SONG Y. C., KAUTZ H., BIGHAM J. P.: Real-time crowd labeling for deployable activity recognition. In *Proceedings of the 2013 conference on Computer supported cooperative work* (2013), ACM, pp. 1203–1212. 15

[Mac04] MACEACHREN A. M.: *How Maps Work - Representation, Visualization, and Design.* Guilford Press, 2004. 7

[MBB*11] MARCUS A., BERNSTEIN M. S., BADAR O., KARGER D. R., MADDEN S., MILLER R. C.: Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2011), ACM, pp. 227–236. 4

[MCG*17] MARTIN D., CARPENDALE S., GUPTA N., HOSSFELD T., NADERI B., REDI J., SIAHAAN E., WECHSUNG I.: Understanding the crowd: Ethical and practical matters in the academic use of crowdsourcing. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments* (Cham, 2017), Archambault D., Purchase H., Hoßfeld T., (Eds.), Springer International Publishing, pp. 27–69. 3, 11, 16

[MDF12] MICALLEF L., DRAGICEVIC P., FEKETE J.-D.: Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2536–2545. 3, 5, 8, 10, 11, 12, 14, 16, 23

[MGGF16] MATEJKA J., GLUECK M., GROSSMAN T., FITZMAURICE G.: The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 5421–5432. 10

[MHRL*17] MCKENNA S., HENRY RICHE N., LEE B., BOY J., MEYER M.: Visual narrative flow: Exploring factors shaping data visualization story reading experiences. In *Computer Graphics Forum* (2017), vol. 36, Wiley Online Library, pp. 377–387. 10

[MKP12] MCDUFF D., KALIOUBY R., PICARD R. W.: Crowdsourcing facial responses to online videos. *IEEE Transactions on Affective Computing 3*, 4 (2012), 456–468. 16

[MMS*08] MCDONALD D. W., MCCARTHY J. F., SOROCZAK S., NGUYEN D. H., RASHID A. M.: Proactive displays: Supporting awareness in fluid social environments. *ACM Transactions on Computer-Human Interaction (TOCHI) 14*, 4 (2008), 16. 4

[MPOW17] MICALLEF L., PALMAS G., OULASVIRTA A., WEINKAUF T.: Towards perceptual optimization of the visual design of scatterplots. *IEEE Transactions on Visualization and Computer Graphics 23*, 6 (2017), 1588–1599. 5, 8, 10, 11, 12, 14, 16

[Mun14] MUNZNER T.: *Visualization Analysis and Design*. A.K. Peters visualization series. A K Peters, 2014. 8, 14

[MVTS15] MUTLU B., VEAS E., TRATTNER C., SABOL V.: Towards a recommender engine for personalized visualizations. In *International Conference on User Modeling, Adaptation, and Personalization* (2015), Springer, pp. 169–182. 17

[NMM*14] NOVAK J., MICHEEL I., MELENHORST M., WIENEKE L., DÃIJRING M., MORÃŞN J. G., PASINI C., TAGLIASACCHI M., FRATERNALI P.: Histograph – a visualization tool for collaborative analysis of networks from historical social multimedia collections. In *2014 18th International Conference on Information Visualisation* (July 2014), pp. 241–250. 16

[OJ15] OKOE M., JIANU R.: Graphunit: Evaluating interactive graph visualizations using crowdsourcing. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library, pp. 451–460. 8, 11, 12

[OMD09] OPPENHEIMER D. M., MEYVIS T., DAVIDENKO N.: Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology 45*, 4 (2009), 867âĂŞ872. 10

[OPH*16] OTTLEY A., PECK E. M., HARRISON L. T., AFERGAN D., ZIEMKIEWICZ C., TAYLOR H. A., HAN P. K., CHANG R.: Improving bayesian reasoning: the effects of phrasing, visualization, and spatial ability. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 529–538. 7, 8, 10

[Pla04] PLAISANT C.: The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces* (2004), ACM, pp. 109–116. 2

[PQMCR17] PADILLA L., QUINAN P. S., MEYER M., CREEM-REGEHR S. H.: Evaluating the impact of binning 2d scalar fields. *IEEE transactions on visualization and computer graphics 23*, 1 (2017), 431–440. 11

[PRS*15] PANDEY A. V., RALL K., SATTERTHWAITE M. L., NOV O., BERTINI E.: How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, pp. 1469–1478. 10

[Pur12] PURCHASE H. C.: *Experimental human-computer interaction: a practical guide with visual examples*. Cambridge University Press, 2012. 7

[PVF14] PERIN C., VUILLEMOT R., FEKETE J.-D.: A table!: Improving temporal navigation in soccer ranking tables. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (2014), ACM, pp. 887–896. 10, 12

[RCFL09] ROBERTSON G., CZERWINSKI M., FISHER D., LEE B.: Selected human factors issues in information visualization. *Reviews of human factors and ergonomics 5*, 1 (2009), 41–81. 2

[RD14] RAHMANIAN B., DAVIS J. G.: User interface design for crowdsourcing systems. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces* (2014), ACM, pp. 405–408. 15

[RELS14] RIEGLER M., EG R., LUX M., SCHICHO M.: Mobile picture guess: A crowdsourced serious game for simulating human perception. In *International Conference on Social Informatics* (2014), Springer, pp. 461–468. 17

[RG15] REINECKE K., GAJOS K. Z.: Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2015), ACM, pp. 1364–1378. 11, 16

[RK11] RZESZOTARSKI J. M., KITTUR A.: Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology* (2011), ACM, pp. 13–22. 16

[RK12] RZESZOTARSKI J., KITTUR A.: Crowdscape: interactively visualizing user behavior and output. In *Proceedings of the 25th annual ACM symposium on User interface software and technology* (2012), ACM, pp. 55–62. 16

[RSC15] RODGERS P., STAPLETON G., CHAPMAN P.: Visualizing sets with linear diagrams. *ACM Transactions on Computer-Human Interaction (TOCHI) 22*, 6 (2015), 27. 8, 10, 12

[RYM*13] REINECKE K., YEH T., MIRATRIX L., MARDIKO R., ZHAO Y., LIU J., GAJOS K. Z.: Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), ACM, pp. 2049–2058. 4, 17

[SBJS15] SARKAR A., BLACKWELL A. F., JAMNIK M., SPOTT M.: Interaction with uncertainty in visualisations. *EuroVis 2015-Short Papers* (2015). 7, 10

[SHC11] SHAW A. D., HORTON J. J., CHEN D. L.: Designing incentives for inexpert human raters. *Proceedings of the 14th ACM International Conference on Computer Supported Cooperative Work (CSCW)* (2011), 275âĂŞ284. 2, 11

[SHK15] SKAU D., HARRISON L., KOSARA R.: An evaluation of the impact of visual embellishments in bar charts. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library, pp. 221–230. 12

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages* (Washington, DC, USA, 1996), VL '96, IEEE Computer Society, pp. 336–343. 7

[SK16] SKAU D., KOSARA R.: Arcs, angles, or areas: Individual data encodings in pie and donut charts. In *Proceedings of the Eurographics / IEEE VGTC Conference on Visualization* (Goslar Germany, Germany, 2016), EuroVis '16, Eurographics Association, pp. 121–130. 8

[SL13] SHESHADRI A., LEASE M.: Square: A benchmark for research on computing crowd consensus. In *First AAAI Conference on Human Computation and Crowdsourcing* (2013). 15

[SM13] SINGER Y., MITTAL M.: Pricing mechanisms for crowdsourcing markets. In *Proceedings of the 22nd international conference on World Wide Web* (2013), ACM, pp. 1157–1166. 2, 11, 15

[SOK*16] STROBELT H., OELKE D., KWON B. C., SCHRECK T., PFISTER H.: Guidelines for effective usage of text highlighting techniques. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 489–498. 7, 11

[SRdSR*15] SOUSA-RODRIGUES D., DE SAMPAYO M. T., RODRIGUES E., GASPAR A. R., GOMES Á.: Crowdsourced clustering of computer generated floor plans. In *International Conference on Cooperative Design, Visualization and Engineering* (2015), Springer, pp. 142–151. 17

[SSG16] SZAFIR D. A., SARIKAYA A., GLEICHER M.: Lightness constancy in surface visualization. *IEEE Transactions on Visualization and Computer Graphics 22*, 9 (2016), 2107–2121. 8, 12

[TBRA17] TURTON T. L., BERRES A. S., ROGERS D. H., AHRENS J.: Etk: An evaluation toolkit for visualization user studies. 4, 16

[TGH12] TALBOT J., GERTH J., HANRAHAN P.: An empirical model of slope ratio comparisons. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2613–2620. 4

[TLCC17] THUDT A., LEE B., CHOE E. K., CARPENDALE S.: Expanding research methods for a realistic understanding of personal visualization. *IEEE computer graphics and applications 37*, 2 (2017), 12–18. 2

[TLM16] TANAHASHI Y., LEAF N., MA K.-L.: A study on designing effective introductory materials for information visualization. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 117–126. 7

[TSA14] TALBOT J., SETLUR V., ANAND A.: Four experiments on the perception of bar charts. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 2152–2160. 12

[TTvE14] TAK S., TOET A., VAN ERP J.: The perception of visual uncertaintyrepresentation by non-experts. *IEEE Transactions on Visualization and Computer Graphics 20*, 6 (June 2014), 935–943. 9

[TWS*] TURTON T. L., WARE C., SAMSEL F., ROGERS D. H., BUJACK R., HALE S. A., MCNEILL G., BRIGHT J., LUZ M., LAWONN K., ET AL.: A crowdsourced approach to colormap assessment. 9

[VAJO13] VÄÄTÄJÄ H. K., AHVENAINEN M. J., JAAKOLA M. S., OLSSON T. D.: Exploring augmented reality for user-generated hyperlocal news content. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (2013), ACM, pp. 967–972. 16

[VC17] VERAS R., COLLINS C.: Optimizing hierarchical visualizations with the minimum description length principle. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 631–640. 7, 8, 10, 12

[vHR08] VAN HAM F., ROGOWITZ B.: Perceptual organization in user-generated graph layouts. *Visualization and Computer Graphics, IEEE Transactions on 14*, 6 (Nov 2008), 1333–1339. 3, 4, 10, 15, 16

[WFI11] WANG J., FARIDANI S., IPEIROTIS P.: Estimating the completion time of crowdsourced tasks using survival analysis models. *Crowdsourcing for search and data mining (CSDM 2011) 31* (2011). 15

[WGK10] WARD M., GRINSTEIN G., KEIM D.: *Interactive Data Visualization: Foundations, Techniques, and Applications*. A. K. Peters, Ltd., Natick, MA, USA, 2010. 7

[WGS*13] WILLETT W., GINOSAR S., STEINITZ A., HARTMANN B., AGRAWALA M.: Identifying redundancy and exposing provenance in crowdsourced data analysis. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2198–2206. 10

[WHA12] WILLETT W., HEER J., AGRAWALA M.: Strategies for crowdsourcing social data analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 227–236. 12

[WYF*17] WAUCK H., YEN Y.-C. G., FU W.-T., GERBER E., DOW S. P., BAILEY B. P.: From in the class or in the wild?: Peers provide better design feedback than external crowds. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 5580–5591. 15

[WZT*16] WANG G., ZHANG X., TANG S., ZHENG H., ZHAO B. Y.: Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 225–236. 16

[XADR13] XUE S., AGARWALA A., DORSEY J., RUSHMEIER H.: Learning and applying color styles from feature films. In *Computer Graphics Forum* (2013), vol. 32, Wiley Online Library, pp. 255–264. 4

[XEZ*15] XU P., EHINGER K. A., ZHANG Y., FINKELSTEIN A., KULKARNI S. R., XIAO J.: Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *CoRR abs/1504.06755* (2015). 7

[YCHZ12] YUAN X., CHE L., HU Y., ZHANG X.: Intelligent graph layout using many users' input. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (Dec 2012), 2699–2708. 16

[YKL11] YUEN M.-C., KING I., LEUNG K.-S.: A survey of crowdsourcing systems. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (2011), IEEE, pp. 766–773. 16

[YLZ14] YANG H., LI Y., ZHOU M. X.: Understand usersâĂŹ comprehension and preferences for composing information visualizations. *ACM Transactions on Computer-Human Interaction (TOCHI) 21*, 1 (2014), 6. 8, 10, 12

[ZCY*11] ZIEMKIEWICZ C., CROUSER R. J., YAUILLA A. R., SU S. L., RIBARSKY W., CHANG R.: How locus of control influences compatibility with visualization style. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 81–90. 12

[ZDKK14] ZHU H., DOW S. P., KRAUT R. E., KITTUR A.: Reviewing versus doing: Learning and performance in crowd assessment. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (2014), ACM, pp. 1445–1455. 15

[ZGB*17] ZHAO J., GLUECK M., BRESLAV S., CHEVALIER F., KHAN A.: Annotation graphs: A graph-based visualization for meta-analysis of data based on user-authored annotations. *IEEE transactions on visualization and computer graphics 23*, 1 (2017), 261–270. 5

[ZGWZ14] ZHAO J., GOU L., WANG F., ZHOU M.: Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on* (2014), IEEE, pp. 203–212. 12

[ZK09] ZIEMKIEWICZ C., KOSARA R.: Preconceptions and individual differences in understanding visual metaphors. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 911–918. 7

[ZK10] ZIEMKIEWICZ C., KOSARA R.: Laws of attraction: From perceptual forces to conceptual similarity. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1009–1016. 8

[ZRH12] ZHI J., ROSENBAUM R., HAMANN B.: Progressive parallel coordinates. In *Visualization Symposium, IEEE Pacific(PACIFICVIS)* (2012), vol. 00, pp. 25–32. 8, 10, 12

**Appendix A:** Example Experiment Reporting Form

**Study:** Assessing the Effect of Visualizations on Bayesian Reasoning through Crowdsourcing
**Experiment:** 1 - comparison of visualization techniques including baseline (i.e., classic problem text with no visualization)

| | |
|---|---|
| Design | Type ○ between-subjects ○ within-subjects ⊗ mixed-design, with between for visualizations, within for Bayesian problems |
| | Independent Variables: Visualization ∈ {V0 (baseline), V1, V2, V3, V4, V5, V6}, Bayesian problem ∈ {Mam, Cab, Eco} |
| | Dependent Variables: Bias (subject's - correct answer), Error (absolute of Bias), Exact answer ∈ {0,1}, Time, Confidence ∈ [1..5] |
| | Covariates: subject's Spatial abilities for paper folding task ∈ [0..10] and Numeracy ∈ [0..30] |
| Data | Source: classic Bayesian problems–Mammography [Eddy,1982], Cab [Bar-Hillel,1980], choosing an Economics course [Ajzen,1977] |
| | Specifics: context and quantitative data comprising of the population size, base rate, hit rate and false alarm rate |
| | Variations: the 3 selected Bayesian problems had diverse context and quantitative data |
| Tasks | Elementary: quantitative comparison and estimation Synoptic: Bayesian reasoning Mixed: n/a |
| | Variations: 3 Bayesian problems with diverse context and quantitative data |
| Vis | Type ⊗ static ○ interactive ○ other |
| | Specifics: generated using custom-built, novel algorithms available at `http://www.eulerdiagrams.org/eulerGlyphs` |
| Test Ques | Question asked: Bayesian problem text (replica of source) + Visualization |
| | Response type: two text entry fields X and Y in the form "X out of Y" |
| | Associated data, task, visualization: 3 Bayesian problems × 7 Visualization |
| Partic | Expected number per independent variable: 24 in total: 168 |
| | Requisites: Amazon Mechanical Turk HIT approval rate >= 95% |
| Quality | Attention detectors: 3 catch questions about the 3 Bayesian problems which appeared after the problems were solved |
| | Multiple participation: data collected from participants who carried out the experiment more than once were excluded from analysis |
| | Others: n/a |
| Motiv | Intrinsic: n/a |
| | Extrinsic: $1 for 25 mins |
| Sys | Hardware: desktop computer (no mobile devices) |
| | Software: JavaScript enabled in web browser |
| Procedure | Training: n/a |
| | Dummy questions (not in training): n/a |
| | Test questions number per participant: 3 ordering: one of the 6 possible problem orderings, with 4 participants for each ordering |
| | Quality assurance questions: 3 catch questions about on 3 Bayesian problems |
| | Abilities tests: 6-question objective numeracy test [Brown et al.,2011]; Subjective Numeracy Scale test part 2 [Fagerlin et al.,2007]; spatial abilities Paper Folding Test (VZ-2) part 1 [Ekstrom et al.,1976] |
| | Questionnaires: 1 about ask participants about their demographics and methods or tools (e.g., calculator) used to solve problems |
| | Steps: instructions+consent, 3 test questions, 3 catch questions, numeracy tests, Paper Folding Test, questionnaire |
| Data Collect | Method: each 6 problem orderings × 7 visualization types was a unique Amazon Mechanical Turk 'external HIT' and 4 copies (assignments) of each were uploaded; HITs were reposted until the required valid data was collected |
| | Managed by Amazon Mechanical Turk Time period: Oct 2011 - Mar 2012 |
| | Testing before actual study: controlled pilot study (N=14); testing in Amazon Mechanical Turk sandbox using diverse web browsers |
| Analysis | Average experiment length: 25 mins Number of participants (before filtering): 266 |
| | Excluded participants with low quality data: 98 for other reasons: 0 in total: 98 |
| | Included participants per independent variable: 24 in total: 168 |
| | Included participants' demographics: 41% female; 32 mean age; 47% in USA, 40% India; 45% only Bachelor's degree; 5 color-blind |
| | Statistics methods used: ANOVA, *t*-test, Bonferroni correction for Visualization and time; Kruskal-Wallis for other Visualization analysis; Pearson's correlation for abilities and between independent variables; histograms and boxplots |
| Repro | Experiment Material: data, tasks, stimuli, online experiment source code, MTurk scripts at `http://www.aviz.fr/bayes` |
| | Collected (Anonymized) Data: all collected anonymized data available at `http://www.aviz.fr/bayes` |

**Table 6:** *An example of how the form in Table 5 could be filled, using the first experiment in Micallef et al.'s 2012 paper [MDF12].*